A study of selection and recombination in microorganisms and a place for the mitochondrion among the α-proteobacteria.

A thesis submitted to N.U.I Maynooth for the Degree of

Doctor of Philosophy

Presented by:

David Anthony Fitzpatrick B.Sc.
Department of Biology
National University of Ireland
Maynooth
Co. Kildare
Ireland.

NUI MAYNOOTH
Ollscoil na hÉireann Má Nuad

October 2004

Supervisor: **Dr James McInerney, B.Sc. Ph.D (NUI)**
Head of Department: **Professor Kay Ohlendieck, Dip.Biol., M.Sc. (Konstanz), Ph.D**

Dedicated to my parents

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr. James McInerney for his supervision of this project. His direction, ideas and belief in what I was aiming to do have been invaluable and I cannot thank him enough.

A big thank you to all in the Bioinformatics lab. Thanks for your friendship and support. I would especially like to thank Dr Chris Creevey who gave up a considerable amount of his time to help with any queries or problems I had.

I would especially like to thank my family for all their support. I would not have made it this far without the three of you and I appreciate all you have done for me. Help came in many forms be it financial or emotional and I knew I could always depend on you.

I would also like to thank Nicola's family for always making me feel very welcome in their home on my numerous visits, I can't thank them enough for their support either.

Finally, I would like to thank Nicola. Your support, love and friendship are amazing and I appreciate the sacrifices you have had to make while I completed this project.

Declaration

This thesis has not been submitted in whole or in part, to this, or any other University for any degree, and is, except where stated the original work of the author.

Signed ____________________

David Fitzpatrick

Abbreviations

ω	the ratio of dN:dS
AI-2	autoinducer-2
BLAST	Basic Local Alignment Search Tool
BS	Bootstrap
CC	Character congruence
dN	The number of replacement substitutions that have occurred between two species since they last shared a common ancestor per replacement site
dS	The number of silent substitutions that have occurred between two species since they last shared a common ancestor per silent site
DNA	Deoxyribonucleic acid
F plasmid	Fertility plasmid
GTR model	General time reversible model
HGT	Horizontal gene transfer
LogDet	Logarithmic Determinant
LRTs	Likelihood Ratio Tests
MCMC	Monte Carlo Markov Chain
ML	Maximum Likelihood
MRP	Matrix Representation using Parsimony
MRC	Matrix Representation with Compatibility
MSSA	Most Similar Supertree Analysis method
NJ	Neighbor joining
OM	Outer Membrane
ORFs	Open Reading Frames
OMP	Outer Membrane Protein
PTP test	Permutation tail probability test
RI	Replacement Invariable substitutions
RV	Replacement Variable substitutions
SH test	Shimodaira and Hasegawa test

SI	Silent Invariable substitutions
SSU rRNA	Small Subunt ribosomal Ribonucleic Acid
SV	Silent Variable substitutions
TRs	Transmembrane Regions
WAG model	Whelan and Goldman model

Nucleotide Abbreviations

A	adenine
C	cytosine
G	guanine
T	thymine

Index of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Index of Tables

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Abstract

This study quantified the prevalence of positive Darwinian selection in a number of bacterial genera. This was achieved using an automated pipeline that utilises newly developed maximum likelihood methods and also parsimony based methods. The percentage of genes found to be evolving under the influence of adaptive evolution was found to be similar to results observed in two eukaryote lineages.

The evolutionary history of several membrane associated bacterial genes was also investigated. Presently a number of these genes are in phase I clinical trials in an effort to investigate their suitability as putative vaccine candidates. A number of these genes were shown to exhibit signatures of positive Darwinian section. This finding has serious implications as recent studies have shown that vaccines that target variable epitopes are less successful then those that target highly conserved epitopes. The implications of these finding are discussed.

A phylogeny derived from four *Neisseria* genomes was proposed using character congruence and total congruence. These two methods are at the centre of heated debate as to which best describes the true phylogeny of large datasets. Systematic problems associated with the character congruence approach were investigated and possible solutions discussed.

An analysis of 31 individual and concatenated protein data sets encoded in *Reclinomonas americana* and *Marchantia polymorpha* mitochondrial genomes revealed that based on the available sequence data, the Rickettsiaceae family are the mitochondrion ancestor. A previous study using the same data and similar methodologies concluded that *Rhodospirillum rubrum* came as close to mitochondria as any α-proteobacterium investigated. A robust phylogeny for the α-proteobacteria is also proposed.

Chapter 1 General Introduction

1.1 Introduction

The genome sequencing revolution of the last ten years has given us a large amount of data, which in turn has led to insights in genome evolution and organisation (Castillo-Davis and Hartl 2003). Due to their relatively small genome size and biological importance there is now a wealth of data pertaining to the prokaryotes and in particular bacterial pathogens.

At a time when millions worldwide are dying from diseases such as meningitis, cholera, typhus, diahoreae; and other diseases such as tuberculosis are re-emerging, it is essential that we try to understand what it is that makes humans susceptible to certain bacterial infections. Two broad mechanisms of evolution have been demonstrated to be important in bacterial virulence. These are positive Darwinian selection, a process that retains useful mutations (Sharp 1997) and horizontal gene transfer a process that allows the acquisition of novel genetic material from the surrounding environs (Doolittle 1999a). Quantifying the number and function of genes that have been subjected to either of these mechanisms will lead to important insights into bacterial virulence. The availability of these sequence data also allows research into the possible relationships among the prokaryotes and should in theory explain some of the long standing questions regarding the origin of eukaryote life.

1.2 Prokaryotes Genomics and Comparative analysis

The prokaryotes consist of genetically distinct unicellular organisms. Often a particular physiological trait unifies and distinguishes groups of prokaryotes. Historically prokaryotes were classified based on observed characteristics such as Gram stain, morphology and motility. Molecular sequences are generally more revealing of evolutionary relationships than are classical phenotypes particularly among microorganisms (Woese *et al.* 1990) and today prokaryotes are mainly grouped on a genetic basis derived from comparisons of small subunit ribosomal RNA (SSU rRNA). Analysis of SSU rRNA have led to the splitting the prokaryotes into two distinct domains the Archaea and the Bacteria, the differences between the two groups are profound (Woese *et al.* 1990).

Since the first complete microbial genome was published in 1995 (Fleischmann *et al.* 1995), more than 200 complete genomes have been completely sequenced and published and another 400 microbial genome projects are presently under way (www.genomesonline.org). The significance of the potential information derived from these genomes should not be underestimated. Considerable improvements in sequencing technologies coupled with reduction in the overall costs have led to the point where sequencing of a microbial genome is now almost routine (Nelson 2003). To date, a number of important human pathogens have been completely sequenced such as *Borrelia burdorferi* (Fraser *et al.* 1997), *Rickettsia prowazekii* (Andersson *et al.* 1998), *Neisseria meningitidis* (Parkhill *et al.* 2000; Tettelin *et al.* 2000) and a number of Chlamydia species (Stephens *et al.* 1998; Read *et al.* 2000; Shirai *et al.* 2000). The genome sequences of other re-emerging pathogens such as *Mycobacterium tuberculosis* (Cole *et al.* 1998) have also been published. The genome sequence of non-pathogens, such as a number of thermophiles have also been completed (Klenk *et al.* 1997; Deckert *et al.* 1998). Some of these may represent the deepest branching members of the bacterial lineage and may help to construct an inclusive prokaryote phylogeny.

A typical bacterial genome consists primarily of coding sequence (more than 90%). This compares to higher eukaryotes where it is not unusual to find less than 10% coding sequence per genome (Salzberg *et al.* 1998). Genome analysis has revealed huge variability in genome size and also guanine and cytosine (GC) content, from a low of 29% for the spirochete *B. burgdorferi* (Fraser *et al.* 1997) to a high of 68% for the actinobacterium *M. tuberculosis* (Cole *et al.* 1998). GC content has been shown to affect the usage levels of certain codons. For example, alanine and proline are found at elevated levels in GC rich genomes while in adenine and thymine (AT) rich genomes methionine and aspartic acid are found at elevated levels. Results from all the completed prokaryote genomes show that there are a large number of predicted coding regions with unknown biological function (Fraser *et al.* 2000). Furthermore a large proportion (10-25%) of the predicted genes are unique to a particular species with no match in other genomes (Fraser *et al.* 2000). This number should decrease as the sample size of genomes is enlarged.

The number of genes involved in certain housekeeping functions such as transcription and translation has been found to be quite similar in all chromosomes even when the genomes differ greatly in size. This would imply that there is a basic complement of proteins that are required for particular cellular processes. Genome analysis has provided unique insights into prokaryote virulence and evolution (Wren 2000). For example the downsizing of obligate intracellular parasites has been observed as has the possible origin of the eukaryote mitochondria (Andersson *et al.* 1998).

1.3 Horizontal Gene Transfer

Initially it was hoped that through complete prokaryote genomes light would be shed on bacterial phylogeny and ancestry which is often unknown due to the lack of identifiable synapomorphies between related bacteria (Snel *et al.* 1999). However, it seems that more questions than answers have arisen with the increase in available genomes (Doolittle 1998). Genes seem to be fluid entities in the bacterial world. Prokaryote phylogenies derived from single genes are rarely consistent with each other (Snel *et al.* 1999). The hobo like existence of prokaryote genes has been blamed on horizontal gene transfer (HGT). HGT refers to the exchange of genes between different strains or species (Doolittle 1999b; Boucher *et al.* 2003). The reason for believing in the occurrence of HGT is relatively simple, genes are not found were they are expected to be (Brown 2003). HGT has been linked to the acquisition of drug resistance by benign bacteria (Hacker *et al.* 2003; Woo *et al.* 2003; Piel *et al.* 2004). HGT has also been shown to lead to the acquisition of genes that confer the ability to catabolize certain amino acids that are important virulence factors (Martin *et al.* 1998). Conflicting hypotheses have been advanced regarding the extent of HGT, variously suggesting that it is restricted to certain categories of genes (Jain *et al.* 1999) or that it was more prevalent in early evolution and not so prevalent now (Woese 2000). One thing is certain though, HGT has so far confounded prokaryote phylogenetic reconstruction and more importantly, through novel disease and antibiotic resistance can affect human health. HGT, once solely of interest for practical applications in classical genetics and biotechnology has now become “the substance of evolution” (de la Cruz and Davies 2000). Conversely some studies have shown minimal genetic transfer into certain bacterial species (Dykhuizen and Baranton 2001). An inability to acquire new genes has its advantages in that parasitic elements such as bacteriophages cannot infect these species (Dykhuizen and Baranton 2001).

The debate regarding the importance of HGT is contentious. Some authors believe it to be the dominant force in shaping prokaryote genomes (Jain *et al.* 2003). Others agree that HGT is an important evolutionary force but disagree that its influence is more important than Darwinian evolution (Snel *et al.* 2002). Recent discoveries of potentially massive gene exchanges in sequenced genomes have again shifted the paradigm (Doolittle 1999a). The resurgence in the pro HGT debate is based on the observations of deviant nucleotide composition and variation of gene content between closely related organisms (Philippe and Douady 2003). Lawerence and Ochman (1998) have even shown that different strains of *E. coli* can differ in gene content by up to 20%. Furthermore foreign DNA could in theory replace a complete genome in only a few hundred million years (Jain *et al.* 2002). Based on this assumption Doolittle (1999b) states that “the history of life cannot be properly represented as a tree” instead he proposes a network type of phylogeny where species are free to swap genes (Figure 1.1). de la Cruz and Davies (2000) express the opinion that the movement of genes via HGT is analogous to “genes flowing in a biosphere as in a global organism”. This radical view concerning a global super-organism is shared by others (Doolittle 2000b; Sonea and Mathieu 2001). Kurland *et al* (2003) argue that there is no doubt that HGT is an important evolutionary force. However they disagree that “HGT is the essence of genome phylogeny” and suggest that the reason for its inflated role is due to the failure to distinguish it from other

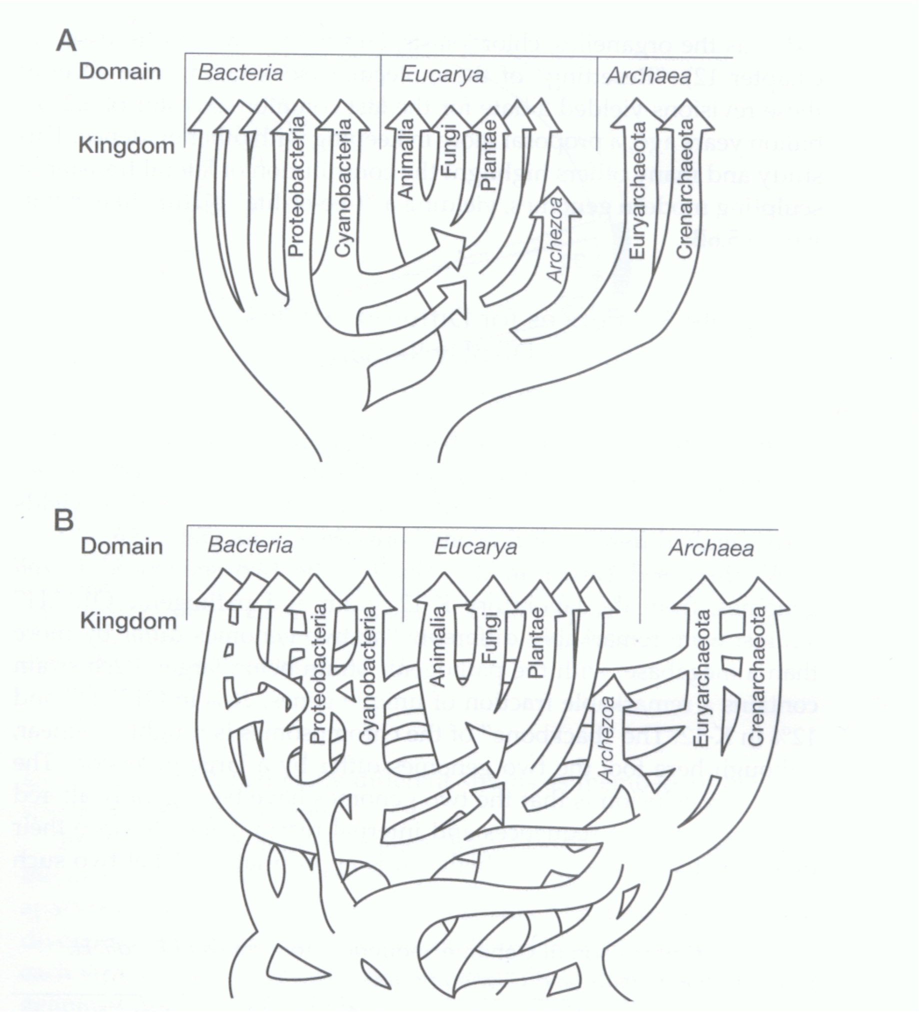


Figure 1.1: Two phylogenetic trees. (A) An orderly tree indicating the origin of the Bacteria, Eucarya and Archaea. In this scenario, the transfer of genes between Kingdom and domains is minimized. The transfer of the mitochondrion from the Proteobacteria and chloroplast from the cyanobacteria into the Eucarya is the only transfer between kingdoms. (B) A highly disorganized tree. It is impossible to locate the origin of the three domains and is the consequence of horizontal gene transfer. Taken from Doolittle (1999b).

phylogenetic anomalies. Gogarten *et al* (2002) showed 30 instances of putative gene transfer. Is this list of 30 genes indisputable evidence of rampant HGT? The list is small when we consider that it is only a very small fraction of the total number of genes to be found in the 30 genomes that were examined. Using 50 complete genomes other researchers have reconstructed a phylogeny that is very similar to that of the SSU rRNA phylogeny (Snel *et al.* 1999). Using similar data but different methodologies other studies have arrived at a similar phylogeny (Brown *et al.* 2001; Korbel *et al.* 2002). The congruence of these phylogenetic reconstructions with the SSU rRNA phylogeny suggests that Darwinian descent is the dominant mode of genome evolution for the genomes used (Kurland *et al.* 2003). There is no doubt that HGT plays an important role in bacterial evolution. The significance of the role it has to play is a contentious one. This can only be solved by analyzing prokaryote genomes as they become available in an effort to fully quantify the major evolutionary forces acting on these organisms. In bacteria there are three broad mechanisms that facilitate DNA movement between cells, these mechanisms are transduction, conjugation and transformation.

1.3.1 Transformation

Natural genetic transformation is the active uptake of free DNA by bacterial cells and the heritable incorporation of this foreign genetic information (Claverys and Martin 2003). The term transformation was first coined by Fred Griffith (1923). Griffith observed that heat-killed encapsulated pneumococci could transfer the ability to make a capsule and hence, to infect mice, when injected together with live, unencapsulated and non-pathogenic pneumococci (Griffith 1923). Since it's discovery, natural genetic transformation has been reported in more than 40 different species (Lorenz and Wackernagel 1994) and provides a potential mechanism for intra and interspecies transfer (Claverys and Martin 2003).

Transformation pathways in Gram positive and Gram negative bacteria are very similar; the main difference being that in gram negative bacteria, DNA must pass through the cell wall and cytoplasmic membrane as well as the outer membrane (Dubnau 1999). Efficient

DNA uptake usually requires uptake sequences that are inverted repeats between genes, acting as transcriptional terminators (Elkins *et al.* 1991). A recent study showed that genes involved in genome maintenance have a higher proportion of these repeats than those genes involved in other functions. Therefore maintenance genes have a higher probability of uptake into a novel strain. This provides a mechanism for facilitated recovery from DNA damage after genotoxic stress for the recepient genome (Davidsen *et al.* 2004).

The initial step in transformation requires binding of double stranded DNA to the bacterial cell surface via active binding sites. The number of active sites varies between different bacterial species,. In *H. infleunzae* the number of active sites for DNA uptake is 4-8 per competent cell (Deich and Smith 1980), this compares with *Bacillus subtillis* where it is estimated that there are approximately 50 binding sites per competent cell (Dubnau 1997). The transformation process next requires fragmentation of the bound DNA by DNAse. Soon after binding, double-stranded DNA fragments are recovered that have been produced by limited cleavage of the initially bound donor molecules (Arwert and Venema 1973). Finally the DNA must be transported into the cell in a 3’-5’ direction (Barany *et al.* 1983) across the cell membrane and wall and results in single-strand integration as the other strand is degraded during transport across the inner membrane (Kahn and Smith 1984). Free cytoplasmic single-stranded DNA of donor origin has yet to be detected in Gram negative bacteria indicating that the incoming DNA searches out a complementary sequence in the recipient (Dubnau 1999).

Transformation requires more than a dozen proteins and is often exquisitely regulated (Dubnau 1997). So what are the selective forces that have shaped and maintained this elaborate system? The obvious theory is genetic diversity through recombination or HGT. Other theories have proposed that transformation has evolved to permit the uptake of DNA as a food supply (Redfield 1993). Another field of thought proposes that transformation serves a function in DNA repair (Wojciechowski *et al.* 1989; Hoelzer and Michod 1991), whereby lysed cells provide DNA that is taken up and used for the repair of otherwise lethal lesions (Love *et al.* 1985).

1.3.2 Conjugation

Conjugation is the directed transfer of DNA from one cell to another usually by conjugative plasmids (Davison 1999). These plasmids contain sequences not only for their replication but also the machinery for transfer between cells via conjugation. The F (Fertility) plasmid of *E. coli* provides an example, this plasmid devotes one-third of its genome to conjugation (Davison 1999). Donor cells that bear pili are designated F^+ while recipients who lack the plasmid are termed F^-. Plasmid transfer between cells is initiated when donor and recipient cells come into contact, the tips of the pili bind to the surface of the F^- cells. Next the pili shorten bringing the two cells close together. Transfer begins from a specific sequence on the donor F plasmid named oriT. DNA is cleaved at oriT by an enzyme relaxase (Byrd and Matson 1997) and transferred to the recipient cell. Transfer is terminated when a complete single copy of the F plasmid has been transferred. Replication proteins in the recipient cell synthesise the complementary strand yielding a double-stranded copy of the plasmid in the recipient cell (Lanka and Wilkins 1995). The original F^- cell is now phenotypically F^+ and can act as a donor cell in subsequent mating.

Genetic transposition is a form of plasmid-independent conjugation. This process involves the relocation of DNA by transposable elements or transposons. In the strictest sense, transposons do not direct the transfer of DNA between cells only the rearrangement of DNA within cells (Berg and Howe 1989). Transposons are important mechanisms in the transfer of antibiotic resistance as they can direct resistance genes into mobile elements such as conjugative elements, which can then relocate these genes to new cells (Bushman 2001). Transposable elements include insertion elements that encode at least one protein that carries out the initial DNA breaking and joining reactions (Berg and Howe 1989). Two transposable elements located close together are known as composite transposable elements. These can transfer all the genetic material between them and have been shown to transfer antibiotic resistance cassettes, as is the case with Tn10, a transposon of *E. coli*. The central region between the insertion elements contains genes for resistance to tetracycline. Some transposons display an ability to transpose to specific target sites that promote the dispersal of the element. For example Tn7 of *E. coli* encodes two genes for antibiotic resistance dhfr and aadA (Bainton *et al.* 1991; Bainton *et al.* 1993). Tn7 transposes at a high frequency to a site in bacterial chromosomes known as attTn7. Transposition into this site does not disrupt nearby genes that encode cell wall biosynthesis and therefore does not debilitate the host cell (Bushman 2001).

1.3.3 Transduction

Transduction refers to the transfer of DNA between two cells using a virus (bacteriophage). Mechanisms of transduction can take many different forms and are generally a normal part of the viral life cycle although they may also occur in error as a secondary by-product of the replication mechanism.

Generalized transduction is a mechanism of transfer whereby any gene from anywhere in a genome can be transferred between cells (Mise and Nakaya 1977). This mechanism of transduction is usually the result of phages that package DNA in their head structure and have accidentally acquired host cell DNA. Infection of new cells by the modified phage can lead to foreign DNA being incorporated into the new host's genome. Phage Mu is an example of a bacteriophage that can lead to generalized transduction (Teifel and Schmieger 1979). Instead of binding to the correct packaging site, the phage DNA packaging machinery binds to host DNA and packets this into the virus head. Transduction of *E. coli* genes has been shown to occur once every 10^7 phage Mu infections (Howe 1973).

Specialized transduction occurs when host genes flanking an attachment site of an integrated prophage are transferred (Bushman 2001) and is the result of the normal excision mechanisms misfiring (Shimada *et al.* 1972). Specialised transduction in *E. coli* K12 has been reported and after thorough studies it was found to be the result of the λ phage (Shimada *et al.* 1973). The gal genes,which are a set of genes responsible for the metabolism of galactose and also bio genes which direct the synthesis of the vitamin biotin have been relocated to newly infected cells (Shimada *et al.* 1973). These genes are in direct contact with the attB region that is the integration site for the λ chromosome (Shimada *et al.* 1973).

1.3.4 Prokaryote Phylogeny

The 16S rRNA gene has been sequenced for a large range of organisms. Phylogenetic analysis of the sequence of this gene has now become the standard way to classify bacteria (Woese 1987; Olsen *et al.* 1994). Because of its universal presence and relatively slow evolutionary rate it can provide information regarding the deep level relationships between the major groups of bacteria. The 16S rRNA gene is only a single gene and may not accurately represent all relationships accurately. In fact Charles Darwin envisaged the problems of classification based on a single character when he said “a phylogeny based on any one character, no matter how important that character may be has always failed” (Darwin 1859).

There is now a large body of evidence for HGT between distantly related species of bacteria (Munoz *et al.* 1998; Brochier *et al.* 2004; Futterer *et al.* 2004). If one accepts that HGT of all gene types is common then we can only arrive at the conclusion that the phylogenetic tree of bacteria is not a tree at all, but rather a tangle of interconnected branches (Doolittle 2000a). The 16S rRNA tree is just one of these trees and in reality may be no more important than any other tree. This may not be the case as recent studies have indicated that there may be a set of genes that are so essential to cell viability that they are in fact recalcitrant to transfer (Jain *et al.* 1999). If there is a core of nonpromiscuous bacterial genes we would expect them to generally agree on the same tree topology and that this tree would truly reflect the evolution of the lineages sampled. It has yet to be established definitively whether such a core of genes exists but there is evidence to suggest that this may be the case (Daubin *et al.* 2002).

The father of the 16s rRNA tree Carl Woese defends it as being a valid representation of the organismal genealogy (Woese 2000). Woese states that HGT may have been common in the earliest organisms but becomes less important over evolutionary time and becomes more confined to closely related species. A recent study constructed phylogenies for deep divisions of prokaryotes and more recent relationships among the γ-proteobacteria (Creevey *et al* 2004). The conclusion reached was that there is almost perfect phylogenetic signal among the γ-proteobacteria but phylogenetic signal within the deepest divisions was no better than random. This finding led these authors to conclude "that deep level relationships are difficult to infer due to extensive HGT, hidden paralogy or the extreme difficulties in inferring deep level phylogenies" (Creevey *et al* 2004). These findings lend support to Woese's argument.

1.3.5 Exploring horizontal gene transfer

As already stated, recent analyses have highlighted high levels of HGT in prokaryotes. The most convincing among these are based on phylogenetic inference (Ragan 2001). Highly supported topological disagreement (incongruence) between trees inferred for one gene family and that inferred for another can often be parsimoniously explained only by invoking HGT (Brown and Doolittle 1997; Nesbo *et al.* 2001). Phylogenetic reconstruction methods remain the only way to reliably infer historical events from gene sequences (Eisen 2000). There are other methods used to infer HGT (discussed below) but phylogenetic methods are the only ones that are based on a large body of work. For example, these phylogenetic methods are designed to accommodate variation in evolutionary rates and patterns within and between taxa. It is not easy to extend phylogenetic methods to all genes, for example some gene families have restricted phyletic distribution and in other cases some genes accept changes so rapidly that orthologs cannot be confidently identified (Ragan 2001). These problems result in sparse trees with weak topological support. Other problems that arise are the computational difficulties of inferring trees and assessing confidence intervals for large data sets. "Unknown optimality for these large datasets is not a promising foundation from which to assess the prevalence of HGT" (Ragan 2001). It is not surprising therefore that there is considerable interest in developing methods that can identify HGT without the need to infer phylogenetic trees. These methods have been referred to as surrogate methods

(Ragan 2001). Examples of surrogate methods include the examination of the patterns of best matches to different species using similarity search techniques to determine the best match for each gene in a genome. This approach has the advantage of speed and automation but does not have a high degree of accuracy. Some notable flaws were brought to attention when the initial publication of the human genome (Lander *et al.* 2001) reported that there are 223 genes that have been transferred from bacterial pathogens to humans. These findings were based on top hits from a basic local alignment search (BLAST) (Altschul *et al.* 1997). Using a phylogenetic analysis these initial claims have been shown to be unsupportable (Salzberg *et al.* 2001; Stanhope *et al.* 2001). Similarly, another study based again on BLAST searches reported that *Mycobacterium tuberculosis* has 19 genes that originate from various eukaryotes (Gamieldien *et al.* 2002), again using a phylogenetic analysis this hypothesis was shown to be unsupportable (Kinsella and McInerney 2003). Reasons for low levels of accuracy with these similarity searches include hidden paralogy, distant slowly evolving genes being detected as best matches or two closely related genes not matching well if they have evolved rapidly (Eisen 1998).

Another major downfall of similarity searches is that they are significantly biased by genome size. In other words the number of best matches to a particular species is dependent in part on the total number of open reding frames in that species and not just evolutionary relatedness (Tettelin *et al.* 2000). A second surrogate method developed by Lawrence and Ochman (1998) identifies regions within genomes that have atypical nucleotide compositions. They reason that when a gene is introduced into a recepient genome it takes time for it to ameliorate to the recipients base composition. Therefore, foreign genes in a genome can be detected by identifying genes with unusual phenotypes such as nucleotide composition or codon usage (Lawrence and Ochman 1997; Lawrence and Ochman 1998). This approach is attractive as it only requires one genome and is not prone to the failings of phylogenetic reconstruction but does suffer from some obvious flaws. For example, atypical composition may be the result of selection or mutation bias. Furthermore, this approach cannot detect transfers between species with similar compositions. Using their own method and the complete sequence of the *E. coli* K-12 genome Lawrence and Ochman (1998) have estimated that approximately 1,600 kb of novel genes have been gained by this strain since diverging from its close relative *Salmonella enterica* over 100 million years ago. In a separate analysis they also reported rates of HGT in 19 bacterial genomes (Ochman *et al.* 2000). Their results (Figure 1.2) show that some genomes such as *Bacillus subtillis* and *Synechocystis* contain relatively large proportions of foreign DNA (12.8% and 16.6% respectively).

Opinion is divided on whether HGT should be estimated using a phylogenetic tree or whether surrogate methods such as those described above can provide robust hypotheses regarding HGT (Ochman *et al.* 2000; Ragan 2001). Surrogate methods do present a heurisistic approach for detecting putative HGT events. A comparison of four surrogate methods by Ragan (2001) highlighted some major problems as all methods fail to identify a common set of genes involved in HGT when analysing *E. coli.* All of the above problems demonstrate the need for a systematic approach to the study of HGT based on first principles that include rigorous inference and statistically based comparison of molecular phylogenetic trees. As more genomic data becomes available this tree-based approach becomes more challenging but only using such an approach can we make definite estimates of rates of bacterial gene transfer.

1.4 Evolutionary Theory

The secrets of heredity and variability in individuals and populations were first explained by Mendel's pea experiments (Mendel 1866). Since Mendel's discoveries biologists have tried to identify the processes that drive evolution. Many hypotheses have been put forward, the most famous of these is Darwin's theory of natural selection.

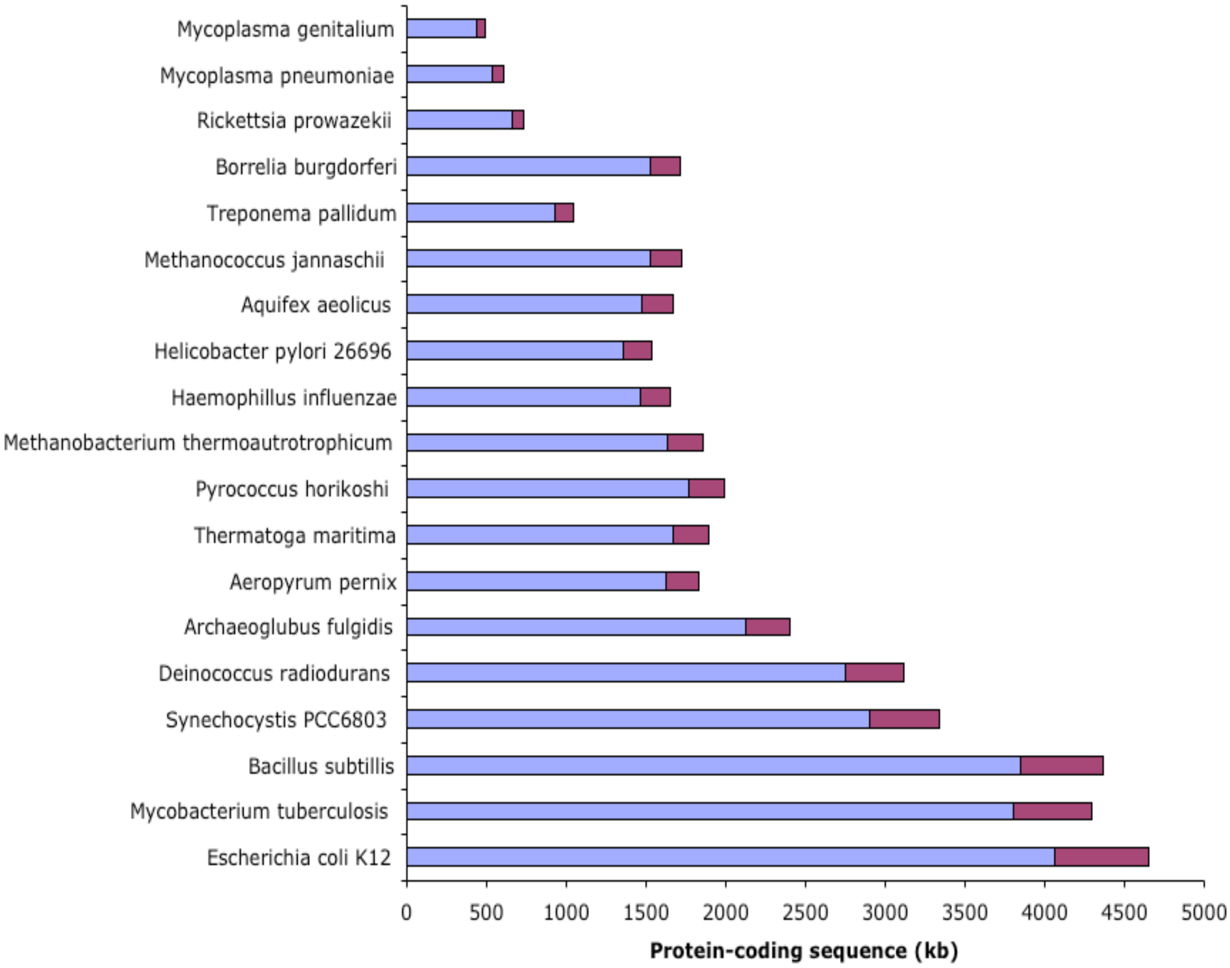


Figure 1.2: Redrawn from Ochman *et al* (2000). Incidences of horizontally acquired DNA in 19 bacterial genomes. Blue bars represent native DNA and pink bars represent foreign DNA.

1.4.1 Natural Selection

After Charles Darwin's voyage to the Galapagos islands on HMS Beagle he proposed that all existing organisms are the modified descendants of one or a few ancestors that we now know arose on earth roughly 3,000 million years ago. He suggested that the main evolutionary force that has led to this change is natural selection (Darwin 1859). The principles of natural selection are as follows. The prime motive for all species is to reproduce, therefore if left unchecked, populations would increase exponentially over time if all individuals survived. Most populations are stable in size however, and a lack of resources to nourish all individuals causes increased competition. Organisms that are most suited to their environment have more chance of survival resulting in the idea of "survival of the fittest". Darwin stated that fitness is passed from generation to generation and subsequent populations of organisms will evolve to suit the environmental conditions. In other words, natural selection predicts the adaptation of organisms to their environments.

Darwin illustrated his theory by documenting the variation in beak sizes of the finches from different Galapagos Islands (Darwin 1859). The Galapagos Islands differ in their natural fauna and consequently each island seems to have its own type of the finch. The finches differ in their beak size and shape depending on the type of food on which they feed. After migration to the islands from mainland South America individual finches that were more suited to obtaining the available resources flourished while those less fit diminished. Genotypes that had a distinct advantage were retained within the population so that today each island has its own similar but distinct species of Finch.

How does natural selection work in an environment that does not change? We would expect that after the fittest genotype is identified it would be preserved by purifying selection. There are documented cases of existing crocodiles being closely associated with reptiles found in the sub-Himalayan fossil deposits (Darwin 1859). There are also fossils of marine *Lingula* (a deep-sea lampshell) that are over five hundred million years old which are identical to their modern day ancestors thus showing the absence of natural selection in constant environmental conditions such as the deep sea after the Palaeozoic era (Darwin 1859).

1.4.2 Synthetic theory of evolution

The late 1960s witnessed a revolution in population genetics. For the first time, protein sequence data was readily available and it was now possible to examine the theories pertaining to the process of gene evolution and substitution. The synthetic theory of evolution (neo-Darwinism framework) was widely championed at this time. This theory recognized mutation as the ultimate source of variation but gave positive selection the dominant role in shaping the genetic makeup of populations. Factors such as mutation and random genetic drift were thought as minor contributors at best; the group of workers who supported this train of thought were called the pan-selectionists. It became obvious however that the level of variability in populations was much higher than expected and not all this variability could be classed as advantageous. This revelation disagreed totally with the synthetic theory of evolution and acted as a spur for new theories of evolution (Hughes 1999).

1.4.3 The Neutral Theory

Kimura's neutral theory of evolution (Kimura 1968) does not deny the role natural selection has to play in shaping the course of evolution but it argues that only a small percentage of DNA changes are adaptive in nature. Instead the majority of evolutionary changes at the molecular level are caused by random drift of selectively neutral mutations (Kimura 1968). The picture of evolutionary change that has emerged from molecular studies seems to be quite incompatible with the expectations of neo-Darwinism (Kimura 1979). For example the rate of amino acid substitution in a given protein is the same in many diverse lineages (Kimura 1968). Similarily the pattern of substitution seems to be random instead of having a pattern and finally the overall rate of change at the DNA level is very high amounting to a least one nucleotide base per genome every two years in mammalian lineages (Kimura 1968).

Evolutionists were quick to criticize the neutral theory; this is not surprising considering the “baroque intellectual climate of the time” (Hughes 1999). However strong support soon came from other work (King and Jukes 1969). These authors arrived at the same result as Kimura independently using molecular data. The fact that synonymous substitutions (those that do not alter the amino acid) are observed far more frequently than non-synonymous substitutions (changes that do alter the encoded amino acid) (Nichols and Yanofsky 1979), provided more supporting evidence for the neutral theory. Neutral mutations are fixed by drift but have no effect on phenotype, and hence none on the individual in the population (Kimura 1983).

Pan-selectionists may have found the neutral theory puzzling but in fact it actually gave them a null hypothesis, which had been absent up to this point (Ohta 1993). The lack of a null hypothesis has a demonstrably detrimental effect on attempts to study Darwinian evolution (Hughes 1999). Gould and Lewontin (Gould and Lewontin 1979) observed that if data were found to contradict one selectionist hypothesis, the tendency was merely to concoct another “adaptive story”. For example, studies of protein polymorphisms of marine vertebrates from tropical, temperate and polar environments have been studied (Mitton 1997). Selectionists predicted that heterozygosity should increase as one moves away from the tropics because environmental variability is greatest in the polar regions and least in the tropics. The exact opposite was observed with greatest heterozygosity found in the tropics. The obvious neutralist explanation would suggest tropical species have larger effective population sizes because of their stable environments and can therefore maintain high levels of neutral polymorphisms; conversely, polar species have lower population sizes because of the variability of their environment and cannot maintain high levels of polymorphism. Instead of the neutralist explanation Mitton (1997) suggests that polar environments forced “species into a strategy of generalism” thus illustrating how different scenarios are conceived to fit the observations made.

The neutral theory assumes that some mutations are deleterious and are therefore eliminated while the remainder are selectively neutral and are maintained (Kimura 1968). The nearly neutral theory or deleterious theory (Ohta and Kimura 1971) makes the same assumptions, as the neutral theory but believes that there should be an additional class of nearly neutral substitutions. This class contains substitutions that are caused by random fixation of very slightly deleterious mutations. Ohta (1995) has presented evidence from 49 single copy genes. The patterns of synonymous and non-synonymous substitutions confirm that the rate of non-synonymous substitutions relative to synonymous substitutions is negatively correlated with the species population size and proves that both random genetic drift and selection are important. This finding is in direct agreement with the nearly neutral theory (Ohta 1995).

1.5 Detecting Positive Darwinian Selection

Positive Darwinian selection or adaptive evolution, is a process that encourages the retention of mutations that are beneficial to an individual or population (Creevey and McInerney 2002). Adaptive evolution has been shown to be responsible for new enzymatic function (Zhang *et al.* 2002), alteration of parasitic membrane exposed surfaces that allows for evasion of host immune responses (Smith *et al.* 1995; Jiggins *et al.* 2002; Urwin *et al.* 2002), alteration of sex-related genes that results in speciation (Lee *et al.* 1995; Swanson *et al.* 2001) and also proteins essential in immune responses (Maurice *et al.* 2001).

Understanding and quantifying levels of adaptive evolution in pathogens could prove useful in future vaccine design. If positive selection is the driving evolutionary force of a pathogen it is theoretically possible to predict what the pathogen should look like in future years. For example a retrospective study that looked at the haemagglutinin gene of influenza concluded that it is possible to predict its evolution over time (Bush *et al.* 1999).

1.5.1 Testing the neutral mutation hypothesis

Polymorphism is a transient phase of molecular evolution, and the rate of evolution is correlated with the level of within-population variation (Ohta and Kimura 1971). One may therefore test the neutral mutation hypothesis by comparing the degree of DNA sequence variation between populations. Many such tests have been developed (Kreitman and Aguade 1986; Sawyer and Hartl 1992).

One test of deviation from neutrality is a method proposed by McDonald and Kreitman (1991). Assuming that there is no recombination, consider a set of alleles from more than one species. The alleles are connected by a phylogenetic tree, which can be divided into two parts: between-species branches and within-species branches. A mutation on a between-species branch will appear in all descendant alleles and is termed a fixed difference between species, whereas a mutation on a within-species branch will be a polymorphism within a species (McDonald and Kreitman 1991). All other sites are monomorphic and are not used in the analysis. Polymorphic and fixed site differences are further divided into two categories, synonymous and non-synonymous. The McDonald and Kreitman method uses a 2 x 2 contingency table to test the independence of one classification (polymorphic versus fixed) from the other (synonymous versus non-synonymous). The test is based on the assumptions that synonymous substitutions are always neutral, only non-synonymous substitutions may be adaptive and a selectively advantageous mutation will be fixed in the population much more rapidly than a neutral mutation therefore it will be less likely to be found in a polymorphic state (McDonald and Kreitman 1991). McDonald and Kreitman demonstrated their model by analyzing the *adh* locus in Drosophila. A consensus sequence was constructed using 30 sequences from three species of Drosophila. When they looked at the substitution types they found a ratio of 7:2 of fixed to polymorphic non-synonymous substitution compared to a ratio of 17:42 of fixed to polymorphic synonymous substitutions. These ratios are significantly different and led to the conclusion that the adh locus has undergone adaptive evolution. Use of the McDonald and Kreitman method on other data sets has revealed positive directional selection at several nuclear loci in Drosophilia (Tsaur *et al.* 1998) and absence of selection at others (King 1998).

A major restriction of the McDonald and Kreitman test is that it can only be used for very closely related species (Sharp 1997). If two genes share a common ancestor they should have a small number of synonymous substitutions this implies that if a significant number of non-synonymous changes have occurred then this facilitates the chance of detecting an adaptive event. If the two genes have diverged a long time ago the number of synonymous substitutions should be greater therefore the ratios calculated by the test will be altered by this saturation of synonymous substitutions. Clearly these tests can only detect adaptation that has happened recently (Sharp 1997).

1.5.2 Pair-wise analysis of sequences

Pair-wise sequence analysis of sequences provides an intuitive method for detecting positive selection. Over the past two decades about a dozen *ad-hoc* methods stemming from pair-wise analysis have been developed e.g. (Nei and Gojobori 1986; Li 1993).

Pair-wise methods differ in the assumptions they make, however they all involve three steps (Ina 1995). Firstly all synonymous and non-synonymous sites are counted between sequences. This step can be complicated by base frequency bias and transition: transversion rate bias. Next the number of synonymous (dS) and non-synonymous (dN) differences between the sequences is determined. This step is straightforward if amino acids only differ at one codon position however they may differ at two or three positions and when this is the case four or six pathways from one codon to another may exist. Ideally these multiple pathways should be weighted appropriately but most methods use equal weightings. Finally, a correction step for multiple substitutions at the same site is applied. Some methods may use Jukes and Cantor's (1969) one parameter model as a mutation matrix while others may use Kimura's (Kimura 1980) two-parameter model. When the rate of dN is greater than dS we can postulate that positive Darwinian selection has occurred.

There are some fundamental problems with some of the methods that have been proposed. Miyata and Yasunaga (1980) and the subsequent simplified version of Nei and Gojobori (1986) both use the Jukes and Cantor (1969) model. The base bias and transition:transversion ratio are both ignored using these models. The problem with these omissions is that they may result in an overestimation of the number of non-synonymous sites, and underestimation of the number of synonymous sites as transitions are more likely than transversions at third positions. A simulation study performed by Ina (1995) showed that of the five methods he tested all gave biased estimates of substitution numbers when there are strong base biases present (such as those found in mitochondrial genes). A study of synonymous substitution rates in Drosophila mitochondrial genes concluded that pair-wise methods should take base bias into consideration (Moriyama and Powell 1997). Another problem arises when we consider that even when we have accurate estimates for dN and dS we may still not be able to detect adaptive events. The dN/dS ratio is averaged over the entire sequence so even if positive selection is acting on a small number of sites in a loop for example, we may loose power in detecting such selection as the number of synonymous substitutions over the entire gene will likely be greater than the rate of non-synonymous substitutions. For example pairwise methods have failed to detect positive selection in the *nef* gene of HIV but using alternative methods such as Bayesian probabilities it has been suggested that certain sites within this gene are evolving positively (Zanotto *et al.* 1999). Similarly certain amino acid sites of a meningococcal surface antigen that are exposed to the human immune system have been shown to be evolving positively again pairwise methods failed to detect any such selection (Urwin *et al.* 2002).

1.5.3 Maximum likelihood methods

A maximum likelihood method to estimate dS and dN between two sequences was developed by Goldman and Yang (1994) based on an explicit model of codon substitution (Goldman and Yang 1994). Maximum likelihood estimation does not involve *ad hoc* approximations and is flexible in that knowledge of the substitution process such as transition:transversion bias, codon usage biases and even chemical differences between amino acids can be incorporated into the model (Yang and Nielsen 2000). The log-likelihood function contains all information about parameters in the model and can measure the fit of different models to the data by comparing their log-likelihood values. The method for estimating dS and dN developed by Yang and Nielsen (1994) is described below

Assume there are n codons in the gene of interest and let the data at site h be $x_h = \{x_1, x_2\}$, where $X1$ and $X2$ are codons at that site in the sequence. The probability of observing data Xh at the site h is

$$f(x_h) = \sum_{k=1}^{61} \pi_k p_{kx1}(t1) p_{kx2}(t2)$$

The term in the sum is the probability that the ancestor has codon *k* and the current species have codons *X1* and *X2* at the site. This probability is equal to the prior probability that the ancestor has codon *k*, given by the equilibrium frequency π_k, multiplied by the two transition probabilities. Since the ancestral codon *k* is unknown all probabilities for *k* are summed. The process is time reversible therefore

$$f(x_h) = \sum_{k=1}^{61} \pi_{x1} p_{kx1}(t1) p_{kx2}(t2) = \sum_{k=1}^{61} \pi_{x1k} p_{kx1}(t1) p_{kx2}(t2) = \pi_{x1} p_{x1x2}(t1 + t2)$$

Parameters in the model are the sequence divergence *t*, the transition:transversion ratio *k*, the non-synonymous/synonymous rate ratio ω, these are usually estimated using a numerical hill-climbing algorithm to maximise ℓ. Another parameter is the codon frequencies π_j these are estimated by the observed base/codon frequencies The log-likelihood function is then given by

$$\ell(t,k,\omega) = \sum_{h=1}^{n} \log\{f(x_h)\}$$

The dN and dS rates are defined as functions of parameters *t*, *k*, ω and π_j and their maximum likelihood estimates are functions of maximum likelihood estimates of parameters *t*, *k*, ω and π_j . Maximum likelihood estimation of dN and dS proceeds as follows, sites and substitutions per codon are counted and partitioned into synonymous and non-synonymous categories so we get

$$\rho_S = \sum_{aa_i \neq aa_j} \pi_i q_{ij}$$

$$\rho_N = \sum_{aa_i \neq aa_j} \pi_i q_{ij}$$

which represent the proportions of synonymous and non-synonymous substitutions respectively. Then we calculate the proportions of synonymous and non-synonymous sites, these are represented as ρ_S^1 and ρ_N^1 respectively. We assume three nucleotide sites in a codon therefore the numbers of synonymous and non-synonymous substitutions sites per codon are then $3\rho_S^1$ and $3\rho_N^1$. The numbers of synonymous and non-synonymous substitutions per site are then

$$dS = t\rho_S/(3\rho_S^1)$$

$$dN = t\rho_N/(3\rho_N^1)$$

ω is the ratio of dN/dS therefore

$$\omega = dN/dS = (\rho_N/\rho_S)/(\rho_N^1/\rho_S^1)$$

From these initial maximum likelihood models others have been developed that allow for different levels of heterogeneity in the ω ratio among lineages (Yang 1998). The simplest model (the “one ratio” model) assumes the same ω for all branches while the most general model (the “free-ratio” model) assumes an independent ω ratio for each branch in the phylogeny. There are also other intermediate models many of which can be compared using the log likelihood ratio test (LRT). Twice the log-likelihood differences of two nested models can be compared to a chi square table with degrees of freedom equal to the difference in parameters between the models tested (Yang 1998). Computer simulations have proved LRTs to be very powerful in discriminating between the fit of two models to the data (Anisimova *et al.* 2001). The methods mentioned above have been successful in locating lineages that have undergone a period of positive selection (Yang 1998; Yang 2002).

Models that allow for heterogeneous selection pressures at amino acid sites have also been developed (Yang *et al.* 2000). These models assume that there are different classes of amino acid sites but that we do not know *a priori* which class each site is from. These models can therefore locate site specific (instead of lineage specific positive selection). Many studies have utilized these models to detect potentially important sites (Zanotto *et al.* 1999; Yang 2000; Fares *et al.* 2001; Kinsella *et al.* 2003). Likelihood “branch site models” (Yang and Nielsen 2002) for detecting positive selection along a specific lineage have been developed based on the above heterogenous models. Recently concerns regarding the frequency of type 1 errors in these models have been raised as it was found that they detect positive selection in 20%-70% of cases when positive selection is absent (Zhang 2004). A number of reasons have been proposed for the poor performance of the branch sites models including over simplistic partition of site classes and the assumption that the transition:transversion ratio for synonymous and nonsynonymous sites is equal. As with all maximum likelihood analyses computation time due to optimization of variables can be very great, this can also increase logarithmically as more taxa are added to the analysis.

1.5.4 Distance based methods

Meisser and Stewart (Messier and Stewart 1997) were the pioneers of a method for detecting adaptive evolution based on a phylogenetic tree. Analysing a primate lysozyme dataset they inferred a phylogeny and from this hypothetical ancestral sequences were reconstructed using maximum-parsimony and maximum likelihood methods at each internal node. They limited comparisons of dN and dS to between reconstructed nodes and their daughters. Meisser and Stewart (1997) located two lineages on which positive selection appeared to be operating. These were the lineages leading to the colobine monkeys and the hominid lineage. A subsequent analysis on the same dataset using maximum likelihood methods by Yang (1998) was in general agreement with the results of Messier and Stewart although Yang argued that the rate of nonsynonymous substitution leading to the colobine lineage was elevated, but not significantly higher than the synonymous rate of substitution. By performing all possible pair-wise comparisons between all sequences (both hypothetical and real) Meisser and Stewart concluded that their method provided greater power when trying to pinpoint the time of adaptive evolutionary events. Furthermore this demonstrates the importance of comparing sequences that diverged within appropriate time frames if positive Darwinian selection is to be detected (Messier and Stewart 1997).

1.5.5 Relative rate ratio test

A relative rate ratio test loosely based on the McDonald and Kreitman (1991) test has been proposed by Creevey and McInerney (2002). The method proceeds by constructing a rooted phylogenetic tree for a given data set. Hypothetical ancestral sequences are reconstructed through a method that uses maximum parsimony (Edward and Cavalli-Sforza, 1963), but in principle any method that accurately reconstructs ancestral character states may be used.

At each internal node all the substitutions in the descendant clades are counted and characterized as either replacement-invariable (RI), replacement-variable (RV), silent-invariable (SI) or silent-variable (SV). Replacement-variable (RV) sites are those where a non-synonymous substitution has occurred in the descendant clade and, subsequently this amino acid position changed again. Conversely, a replacement-invariable (RI) site is one where a non-synonymous substitution has occurred at a particular amino acid position but the new amino acid has not changed subsequently. Similar counts for SI and SV are calculated for synonymous substitutions. If the genes are evolving neutrally then the ratios of RI:RV to SI:SV should be similar, this is analogous to the McDonald and Kreitman test (1991) . If positive selection has characterized the evolution of the genes being examined the number of either RI or RV substitutions will be significantly greater than is expected from neutrality. This method has been very effective at identifying adaptive evolution in a number of instances (Creevey and McInerney 2002).

1.5.6 Bayesian posterior probabilities

Locating amino acids that are positively selected is essential if we wish to make a prediction about the function of a protein. An amino acid undergoing positive selection is likely to play a key role in the function of the protein (Nielsen and Huelsenbeck 2002). Bayesian methods for identifying positively selected sites have been proposed (Yang *et al.* 2000; Nielsen and Huelsenbeck 2002) and these methods are an integral part of studies that generate hypotheses for the evolution of proteins in natures evolutionary experiment (Yang and Bielawski 2000). Bayes theorem can tell us the posterior probability of a site having dN/dS > 1. Bayes theorem is given by

$$\Pr ob(H \mid D) = \frac{\Pr ob(H * D)}{\Pr ob(D)}$$

Where H is a given hypothesis and D is a particular dataset. In methods for inferring sites under positive selection, D is the data at a site, and H is the site class with a particular dN/dS ratio (Yang and Bielawski 2000). Any unknown parameters such as branch lengths and parameters in the dN/dS distribution over sites are calculated by averaging the probability of observing data at a site from all sites this constitutes the likelihood. After parameter estimation, the Bayes theorem can be applied to calculate the probability that any site, given data at that site is from a particular site class. Simulation studies have shown Bayes prediction of amino acid sites being under positive selection to be very accurate (Anisimova *et al.* 2002).

Alternative approaches in calculating Bayesian posterior probabilities have been suggested (Nielsen and Huelsenbeck 2002). This new approach explicitly maps mutations onto a phylogeny and instead of estimating parameters using a likelihood function, inferences are based directly on the posterior distribution of mutations. Despite the differences in the two approaches, the biological conclusions from studies using both approaches on the hemagglutinin protein of the influenza virus are remarkably similar (Nielsen and Huelsenbeck 2002).

1.6 Supertrees

Constructing a prokaryotic "Tree of Life" is a goal of microbial systematics. Over the last three decades, microbial phylogenetists have used an ever-expanding database of SSU rRNA to infer prokaryote phylogeny (Doolittle 1999a). Advocates of this approach point to the vertical transmission of this gene as well as its ubiquity and slowly evolving sites (Woese 1987). There are opponents to this approach however who point out that phylogeny reconstruction based on a single gene may not be robust as vital physiological processes (i.e. photosynthesis and methylotrophy) and basic adaptive strategies (i.e. halophily and thermophily) do not map simply to the SSU rRNA tree (Boucher *et al.* 2003).

An alternative approach to a single gene phylogeny is to combine all the available phylogenetic data available (Bininda-Emonds *et al.* 1999). A supertree does this by generating a phylogeny from a set of input (source) trees that posses fully or partially overlapping sets of taxa. Because the source trees need only overlap minimally each source tree must share at least two taxa with one other source tree. The supertree will contain all taxa found in the set of source trees. The above definition makes a distinction between supertrees and consensus methods, which only combine fully overlapping source trees. Consensus methods are described later.

1.6.1 Supertree construction

Supertree techniques can be classified broadly as either “direct” or “indirect” (Wilkinson *et al* 2001). Direct methods are similar to classical consensus techniques whereby the output tree is derived directly from the source trees without an intermediate step. Indirect methods use some form of matrix representation to encode individual source tree topologies as matrices that are combined and analysed using an optimization criterion (Bininda-Emonds *et al* 2002) (Figure 1.3).

The best-known and most widely used supertree construction method is *matrix representation with parsimony* (MRP) (Baum 1992; Ragan 1992). MRP is actually a consensus method (see below) that has been adapted to build supertrees. MRP creates a matrix with characters referring to the topologies of the source trees. The coding scheme

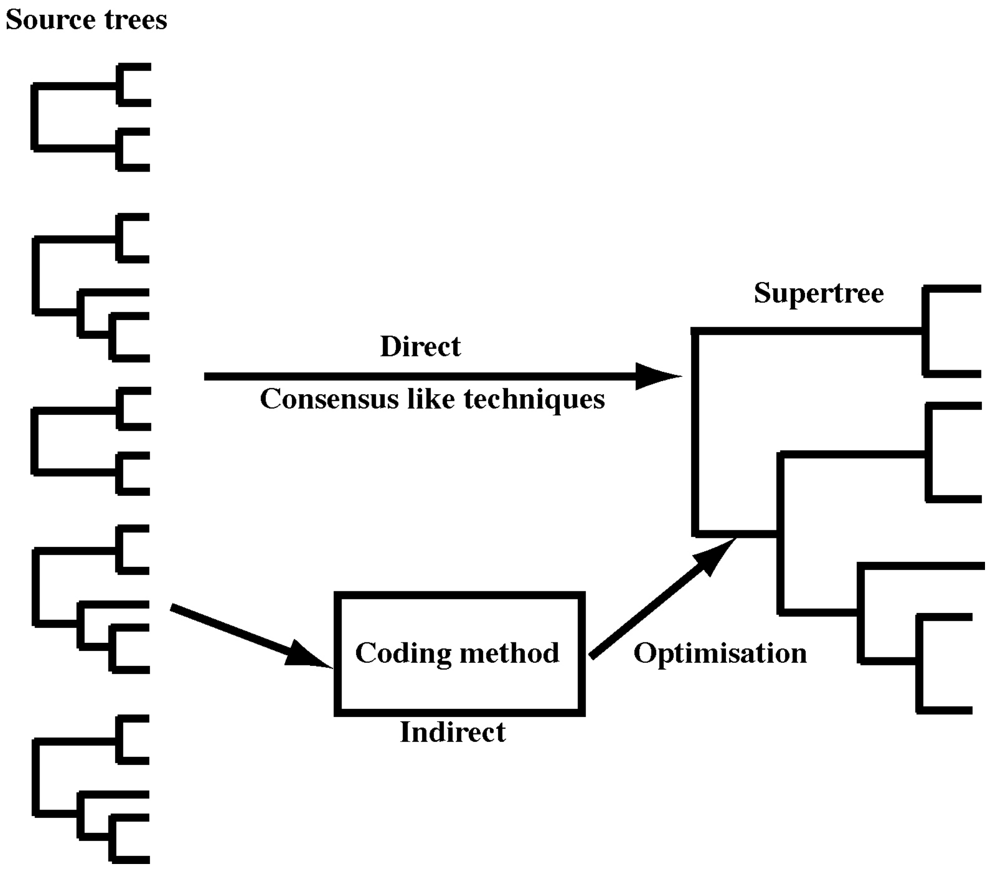
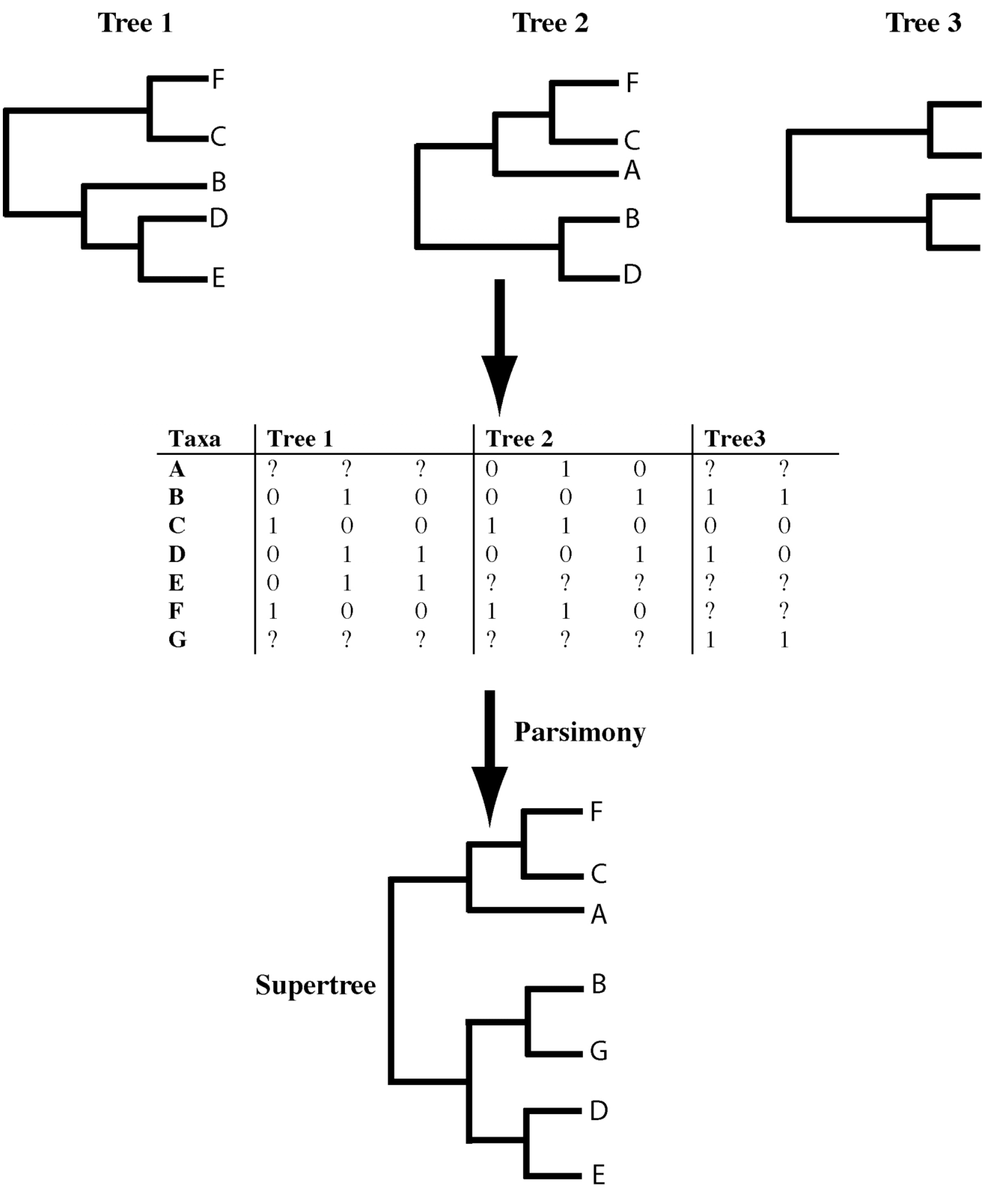


Figure 1.3: Representation of direct and indirect consensus techniques. Direct methods combine all available data into a single tree while indirect methods encode the input trees into a matrix and then an optimality procedure analyses the matrix to produce the end supertree.

scores a ‘1’ for each taxon in a clade and ‘0’ for taxa contained within the source tree but not in the clade. A ‘?’ is assigned to any taxa not present in the source tree (Figure 1.4). The columns in the matrix each represent one internal branch in one of the source trees therefore the total number of columns in the matrix is equal to the total number of internal branches across all the source trees. The matrix is analysed with parsimony (Edward and Cavalli-Sforza, 1963) and the supertree phylogeny is constructed. Although MRP is widely used there are some known problems. For example MRP has a known size bias in which the more inclusive of two competing analogous clades is favoured in the supertree because it contributes more characters to the matrix i.e. “the bigger the tree the bigger the vote” (Bininda-Emonds and Bryant 1998). There are many other supertree methods that utilise the matrix representation described above. These include the *MinFlip* supertree method (Chen *et al* 2003) which flips the contents of individual matrix cells from 0 to 1 or vice versa in a bid to remove conflict, *MinFlip* supertrees equate to the matrices in which conflict has been removed in the least number of flips. Another method that utilises matrix representation is *matrix representation with compatibility* (MRC) (Purvis 1995). MRC finds the largest set of consistent data and uses this to construct a supertree, inconsistent data that is incompatible with the majority is excluded, this step removes the need for parsimony analysis.

There are also a number of supertree methods that use distance matrix representations of trees. Lapointe and Cucumel (1997) have developed the *average consensus* method, this uses least squares (Cavalli-Sforza and Edwards 1967) analysis of the matrix of average pairwise path-length distances. The *average consensus* procedure is the only described method that uses branch lengths when they are available. The *most similar supertree analysis* (MSSA) method (Creevey *et al* 2004) uses matrix representations of equal branch length path distances and is similar in this respect to Lapointe and Cucumel’s average consensus method (1997). MSSA proceeds by proposing a supertree based on all the taxa found within the source trees. Considering each source tree individually the supertree is pruned until both trees possess identical leaf-sets, then a simple tree-to-tree distance is used to evaluate similarity between the pruned supertree and source tree (Figure 1.5), the sum of all the pairwise distances is added and this gives the dissimilarity



Taxa	Tree 1			Tree 2			Tree3	
A	?	?	?	0	1	0	?	?
B	0	1	0	0	0	1	1	1
C	1	0	0	1	1	0	0	0
D	0	1	1	0	0	1	1	0
E	0	1	1	?	?	?	?	?
F	1	0	0	1	1	0	?	?
G	?	?	?	?	?	?	1	1

Figure 1.4: A Supertree of three source trees created using MRP. The trees are encoded into a matrix using the coding scheme proposed by Baum (1992) and Regan (1992) and then parsed by parsimony to give a supertree.

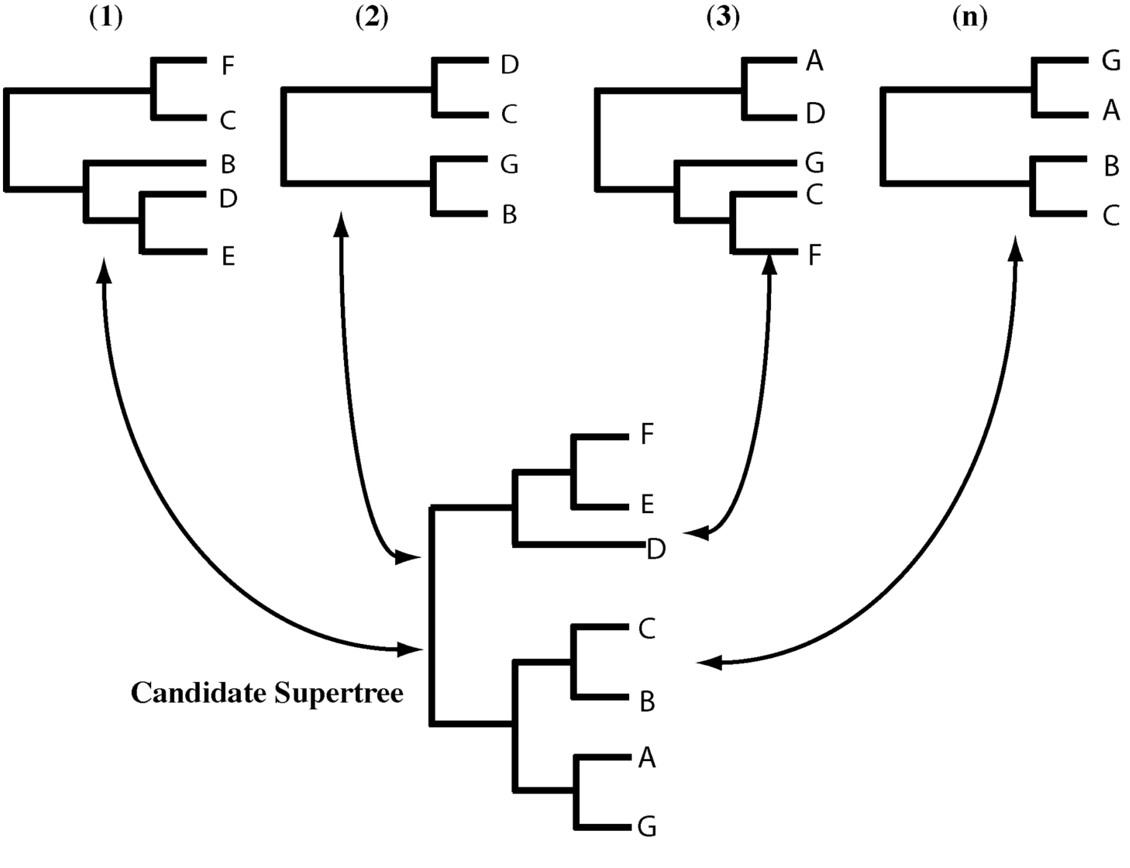


Figure 1.5: Graphical representation of the MSSA method. The similarity X_i of any input tree to the appropriately pruned supertree is calculated by a simple tree-to-tree distance. The sum of these pairwise distances gives the dissimilarity of the trees. To normalise for large tree bias the sum is divided by the total number of comparisons. A supertree is assigned a score of zero if, for all gene trees its subtree on the gene trees leaf set was identical to the gene tree. Higher scores indicate increasing dissimilarity.

of the trees. To normalise for large tree bias the sum is divided by the total number of comparisons. Using this approach a proposed supertree is assigned a score of zero if for all gene (source) trees and their corresponding subtrees are identical (Creevey *et al.* 2004). Higher scores indicate increasing dissimilarity. Numerous other tree-to-tree distance measures could be used to define optimal supertrees. An exhaustive search of all possible supertrees can be carried out to find the optimal supertree. Supertree construction is a relatively new approach for combining phylogenetic information (Bininda-Emonds *et al.* 1999). Furthermore our understanding of supertree methods are limited therefore some authors (Wilkinson *et al* 2004) have warned against an over-reliance upon any single method and instead propose the use of multiple methods, which will allow assessment of each method and any possible biases with respect to input tree shape or size. To date almost all supertrees (those that utilise the largest set of gene families as possible) have been constructed using MRP (Daubin *et al.* 2001; Jones *et al.* 2002; Pisani *et al.* 2002; Ruta *et al.* 2003). Ideally more than one method should be used and the phylogenies created from all methods compared, to limit the likelihood of methodological artefacts.

1.6.2 Consensus methods

Fundamentally, consensus methods are nothing more than a function that returns a single tree (Bryant 2003). For consensus methods, the set of input trees must all contain the same taxa; this is in contrast to the supertreee construction methods described above which can account for missing data. Many consensus methods have developed over the last three decades since “ a new problem in the science of classification… along with its solution” was presented by Adams (1972). The problem Adams addressed was the best way to combine rival trees into a representative tree and the solution was the first consensus method known as the *Adams consensus tree*.

As the *Adams consensus tree* was the first consensus method it is no surprise that it is one of the more popular consensus methods . The method is defined for rooted trees only and cannot handle unrooted trees (Steel *et al.* 2000), it proceeds by finding all three-taxon statements that are not contradicted by any of the input trees and then creates a consensus from them (Figure 1.6c. for explanation). The *strict consensus* method of Rohlf (1982) is the simplest consensus method (Bryant 2003). The strict consensus tree includes only monophyletic groups that occur in all trees being considered or put another way only the groupings (splits) that are common to all trees will be found in the strict consensus tree, all other taxa will be represented as polytomies (Figure1.6a).

One of the more liberal consensus methods is the *majority-rule consensus* (Margush and McMorris 1981). This method constructs a consensus tree that contains all the groups that occur more than a predefined percentage of the time, for example a 50% majority-rule consensus tree will only contain all groups/splits that occur more than 50% of the time (Barthelemy and McMorris 1986, Bryant 2003) see Figure 1.6b for an explanation of the method. Other methods include the Greedy consensus tree, which allows additional clusters to be included in the majority rule tree (Bryant 2003). The validity of using consensus and supertrees methods to create phylogenies (an approach termed taxonomic congruence, see below) is at the heart of a vigorous debate in systematics (Felsenstein 2004), opponents of these methods instead propose a character congruence approach.

1.6.3 Character congruence

Character congruence (CC) was first introduced by Kluge (1989), it proposes that phylogenetic analysis should always be performed using all the available character data. In particular all of the independent characters available to the systematist should be combined (in affect concatenating sequences end to end) and then analysing them using parsimony, although any other phylogenetic method could also be used (Huelsenbeck *et al* 1996). Kluge (1989) justified the character congruence approach on philosophical claims insisting that it maximises the informativeness and explanatory power of the character data used in the analysis. In an ideal case CC should converge to the correct phylogenetic tree as more data is included. Kluge (1989,1998) points out that CC takes many different loci in account and does not depend on an arbitrary choice of a consensus tree method. Using a CC approach, various phylogenies have been constructed recently

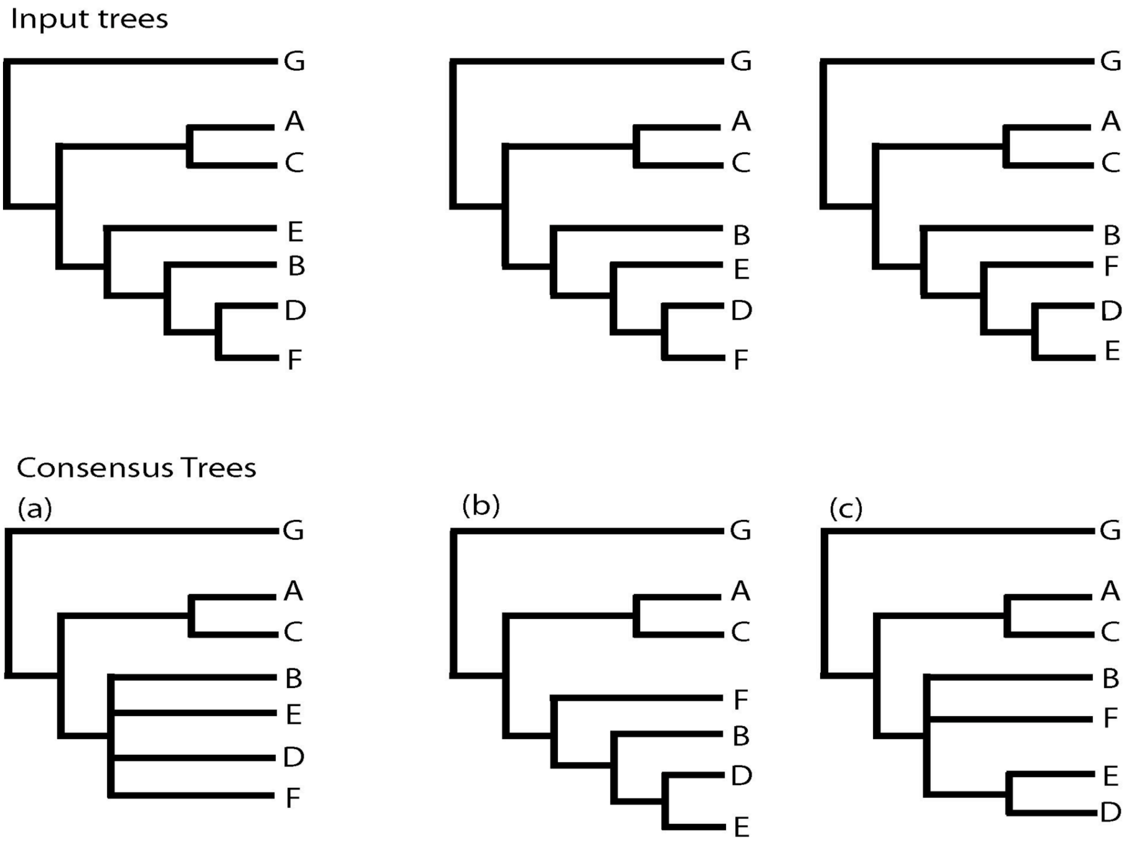


Figure 1.6: A set of 3 input trees all with the same taxa and three consensus trees created using different methods. (a) *Strict consensus*, the grouping of A and C is the only common grouping to all input trees. (b) 50% *majority rule consensus*, this tree only contains the clusters that appear in more than half of the input trees.(c) *Adams consensus* tree of the input trees is formed by finding all of the three taxon statements that are not contradicted by any of the trees. Three taxon statements are triplets of species that show to of the species to be more closely related than the third. In the input trees, the triple species ((D,E),B) appears in all, therefore the Adams consensus will contain the grouping (D,E) showing that these are nested relative to B. Conversely the triplet BEF does not that is common to all trees so F, B and E are represented as a trifurcation.

such as a bacterial phylogeny (Brown *et al.* 2001), an archael phylogeny (Brochier *et al.* 2004) and a yeast phylogeny (Rokas *et al.* 2003).

The alternative approach to CC is termed taxonomic congruence (TC). This method argues that independent datasets should be analysed separately and combined by means of consensus techniques *a posterior* (Bull *et al.* 1993). Some of these techniques include supertree methods and also the original consensus methods described above. Miyamoto and Fitch (1995) have reasoned that by performing separate analyses of partitioned data (genes) the systematist gains important information that would be otherwise unavailable if the data were combined *a priori*. Partitions may include slow and fast evolving genes or partitions based on gene function. It is argued that congruence among different partitions provides strong evidence that a particular phylogenetic estimate is accurate (Penny and Hendy 1986). Furthermore, other authors (Miyamoto and Fitch 1995) have noted that if different loci have substantially different rates of change then combining them into a single data set may obscure evidence that indicated that one locus should be treated differently from another. For example if all but one locus experience a low but equal rate of evolution then the end phylogeny may be directly influenced by the phylogenetic relationships gleaned from a single locus with a high evolutionary rate. Initially numerous studies declared CC to be superior to TC (Barrett *et al* 1991, De Queiroz 1993) as CC usually provides trees that are better resolved than those obtained by TC. More recently, TC methods such as the *average consensus* (Lapointe and Cucumel 1997) have been shown to do better than those based on topological relationships alone (i.e. strict and majority rule) and furthermore may produce trees that are as well resolved as those obtained from TE (Lapointe *et al* 1997).

The controversy regarding CC and TC methods is ongoing. Instead of choosing one approach over an other, one is probably better to follow the suggestion made by De Queiroz (1993) and Larson (1994) where they advocate that both analyses should be performed and then the trees gleaned from both analyses cross-corroborated. This approach has been termed global congruence (Levasseur and Lapointe 2001).

Chapter 2 An analysis of variable selective pressures on the protein-coding components of a variety of bacterial genomes

2.1 Introduction

The sheer quantity of prokaryote genome sequences available today provides a unique opportunity to elucidate bacterial evolution at the molecular level through the use of large-scale genome comparisons. It is widely accepted that the majority of DNA divergence between species and strains is the result of mutation and drift as explained in Kimura's neutral theory (Kimura 1979). However, there is a growing body of evidence to suggest that positive selection i.e. (the fixation of advantageous amino acid substitutions) is also an important evolutionary process, although the extent of its influence has been much debated. To fully appreciate bacterial or any organisms' evolutionary mechanisms, it is important that we can quantify the numbers of genes that have undergone positive selection. Insights into the biological functions and evolutionary mechanisms of these genes will help us to understand how specific organisms react to changes in their environment. In an effort to quantify the prevalence of positive selection, a large-scale search for genes on which positive selection may operate has been conducted in the past (Endo *et al.* 1996). The approach taken by these authors assumes that the number of nonsynonymous substitutions per nonsynonymous site (dN) should be greater than the number of synonymous substitutions per synonymous site (dS) in genes on which positive selection operates. This analysis identified 17 proteins out of 3,595 (0.47%) with a higher than expected nonsynonymous substitution rate. Traditionally positive selection has been defined using these criteria (i.e. a ratio of dN / dS > 1). However, it is well known that these criteria can fail to identify positive selection when this ratio is averaged over the entire protein, making this approach quite weak at detecting positive selection.

In this chapter, I sought to determine the proportion of bacterial genes from various genera that exhibit evidence of having being under the influence of adaptive evolution.

The results of this analysis directly addressed the question regarding the prevalence and possible role positive Darwinian selection plays amongst closely related bacterial strains. The approach taken was a scaled down version to that of Endo *et al* (1996) in that the databases that were used only contain sequences from four closely related bacterial strains. A large body of published work has illustrated the ability of membrane associated proteins to evade host immune responses through positive Darwinian selection (Bush *et al*. 1999; Fares *et al*. 2001; Jiggins *et al*. 2002; Urwin *et al*. 2002; Andrews and Gojobori 2004) therefore this category of gene function was singled out as potentially being of interest. The approach taken in this analysis was similar to that of Endo *et al* (1996) with additional use of recently developed maximum likelihood (ML) methods for testing for evidence of positive selection (Yang *et al*. 2000). The genera examined in this work are listed below; each genus contains at least one human pathogen.

1) The *Neisseria* genus, which contains pathogens that cause septicaemia, meningitis, and gonorrhoea. Possible reasons for successful speciation between *Neisseria* commensals of the urogenitary tract and nasopharynx are also addressed using both a ML and parsimony framework.
2) The *Chlamydia* genus, which contains pathogens that cause several human diseases of medical significance including trachoma a leading cause of preventable blindness and infections of the genital tract often leading to ectopic pregnancy and inflammation of the pelvis.
3) The *Bacillus* genus, which includes pathogens that causes inhalation anthrax.
4) The *Escherichia* genus with strains that can cause food borne illness resulting in diarrhoea and nausea.

The ML framework used, utilises codon-based models that account for variation in dN/dS ratio (ω) (Yang *et al.* 2000), ω is allowed to assume a value from a number of site classes. The application of these models has proved very successful and has led to the detection of positive selection in many genes for which it had not been previously been suggested e.g. (Zanotto *et al.* 1999; Yang *et al.* 2000). However to date no large scale search for positive selection has utilised these ML models. This is probably due to the computational costs associated with these methodologies. A recent study by Anisimova *et* *al* (2001) into the accuracy and power of ML methods has been conducted and has shown ML methods to be very robust when the sample size and sequence length are large enough. Other recent simulation studies have pointed to instances where ML methods falsely identify positive selection (Suzuki and Nei 2001; Suzuki and Nei 2004). The ML method sorts codon sites into different ω categories. Sites with high ω values are grouped into the ω > 1 category making this method efficient at detecting selection if the high ω values are indeed caused by selection (Suzuki and Nei 2004). In reality, ω is affected by stochastic errors, for example it is possible to get an infinite ω value at sites where synonymous sites do not exist (Suzuki and Nei 2004). The size of samples in this analysis are smaller than those recommended by Anisimova *et al* (2001) (i.e. each gene family contains 4 orthologous genes) so simulation studies that address these concerns were performed (Anisimova *et al.* 2001). In an effort to circumvent the potential problems associated with a ML analysis of positive selection, multiple testing of a number of evolutionary models was performed. Strict criteria for the parameters indicative of positive selection were also set. It is hoped that this approach taken minimises many of the pitfalls associated with such an analysis.

2.2 Materials and methods

2.2.1 Databases and searches

Four nucleotide databases were constructed, each contained four strains from a particular genus (Figure 2.1). Nucleotide databases were translated into their amino acid equivalents. The analysis was limited to closely related strains to minimise the affect of saturation of synonymous sites and recombination. Saturation of synonymous sites leads to the underestimation of dS and therefore an inflated dN/dS ratio (Lynn *et al.* 2004), while recombination leads to apparent substitution rate heterogeneity and can closely resemble the effects of molecular adaptation. The four genera examined were

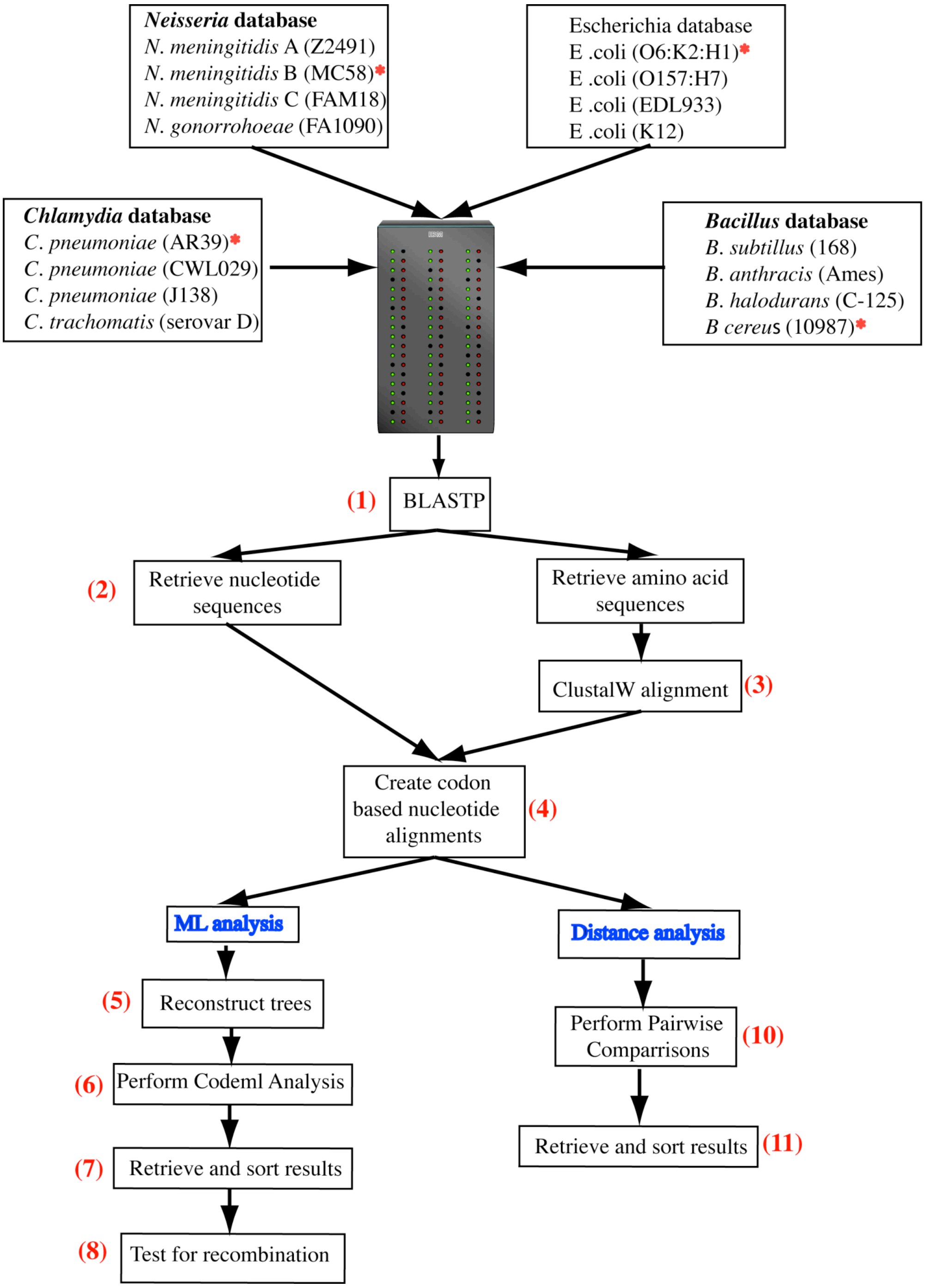

Neisseria database
N. meningitidis A (Z2491)
N. meningitidis B (MC58)*
N. meningitidis C (FAM18)
N. gonorrohoeae (FA1090)
Escherichia database
E .coli (O6:K2:H1)*
E .coli (O157:H7)
E .coli (EDL933)
E .coli (K12)
Chlamydia database
C. pneumoniae (AR39)*
C. pneumoniae (CWL029)
C. pneumoniae (J138)
C. trachomatis (serovar D)
Bacillus database
B. subtillus (168)
B. anthracis (Ames)
B. halodurans (C-125)
B cereus (10987)*
(1) BLASTP
(2) Retrieve nucleotide sequences
Retrieve amino acid sequences
ClustalW alignment (3)
Create codon based nucleotide alignments (4)
ML analysis
Distance analysis
(5) Reconstruct trees
(6) Perform Codeml Analysis
(7) Retrieve and sort results
(8) Test for recombination
Perform Pairwise Comparrisons (10)
Retrieve and sort results (11)

Figure 2.1: Automated pipeline for detecting positive selection. Each database consisted of four species from a single genus. Strains with red stars beside them were used are the query genome for database searches.

Escherichia, *Bacillus*, *Chlamydia* and *Neisseria* (Figure 2.1). For the *Neisseria* database *N. meningitidis* serogroup C (FAM18) and *N. gonorrhoeae* (FA1090) have both been sequenced but as of yet are unannotated. The complete genomes of both strains were provided by the Sanger centre (ftp://ftp.sanger.ac.uk/pub/pathogens/nm) and the University of of Oklahoma (ftp://ftp.genome.ou.edu/pub/gono) respectively.

Potential coding regions for both *N. meningitidis* C and *N. gonorrhoeae* were located using the microbial gene identification tool Glimmer 2.0 (Delcher *et al.* 1999). Glimmer is a computational gene finder that can find 97-98% of all genes in a prokaryotic genome without any human intervention (Delcher *et al.* 1999). Glimmer uses Markov chains which model the probability of a given base being part of a gene depending on the preceding bases. All putative open reading frames (ORFs) identified were extracted and those that did not overlap with other ORFs were retained. The software was trained to recognise *Neisseria* genes using the genome of *N. meningitidis* serogroup B. This process produces a set of ORFs that are likely to be coding (Delcher *et al.* 1999).

2.2.2 Pipeline for detecting positive selection

As a number of bacterial genera were to be tested, an automated pipeline (Figure 2.1) to detect positive selection was developed. The pipeline uses UNIX shell scripts and a number of novel JAVA programs. The advantages of this approach are that it is repeatable and can be readily used in future when new sequence data becomes available. The pipeline follows procedure listed below.

1) Homologous sequences were identified by performing a database similarity search of the bacterial databases using the BLASTP algorithm (Altschul *et al.* 1997) with a cut off expectation value of 10^{-7}. For each database, one genome within that database was nominated as the query genome, the genome with the largest number of ORFs was generally selected.
2) From the multiple BLASTP searches, single gene families that contained all four strains were retained and their corresponding nucleotide and amino acid sequences extracted from the relevant database. The datasets were strictly limited

to single gene families, as it can be difficult to infer orthology with complete confidence among multigene families.

3) The single gene amino acid sequences were aligned using ClustalW 1.81 (Thompson *et al.* 1994) using the default settings.
4) Gaps created in the amino acid alignments were transposed back to the nucleotide sequences to obtain codon based nucleotide alignments. Positions where stop codons occur were stripped as these cause the ML software to quit.
5) Neighbor joining (Saitou and Nei 1987) trees based on protein distances using the neighbor and protdist programs from the PHYLIP (Felsenstein 1989) package were created.
6) JAVA code kindly provided by Mary O'Connell of the bioinformatics lab of N.U.I. Maynooth was modified to create all the directories and control files for the ML software. A maximum likelihood (ML) approach implemented in the PAML package 3.13 (Yang 1997) was used to examine selection pressures acting on bacterial homologues. Three likelihood ratio tests (LRTs) were used. The first compares model M0 (1ratio), which assumes one ω for all sites with M3 (discrete k=2) that assumes two site classes with independent ω values estimated from the data. The second LRT compares M1 (neutral), which allows two site classes with values fixed at 0 and 1 with M2 (selection) which has an additional site class that allows $\omega > 1$. The final LRT used compares M7 (beta) which allows for 10 site classes (each with $\omega < 1$), with M8 (beta & ω), which has an additional site class that allows $\omega > 1$. To avoid being trapped at a local optimum, it is important to run M3 and M8 at least twice (once with initial $\omega > 1$ and once with $\omega < 1$) and results corresponding to the highest likelihood value should be used (Yang 1997; Anisimova *et al.* 2001). Multiple starting omega values (four in total 0, 0.5, 1.2, 3.14) were selected for models M3 and M8.
7) All results were sorted and the relevant LRTs performed. The different site classes and the proportion of sites within that class were also determined and the end result is a tab-delimited file, which is readable by Microsoft Excel. From the ML analysis of positive selection, strict criteria were set in order for a gene family to be a candidate for being under the influence of positive selection. These criteria

stated that for the three nested LRTs performed (M0 *vs* M2, M1 *vs* M2 and M7 *vs* M8) all must be significant at the 5% level. Furthermore, at least 5% of the sites must have $\omega > 1.3$, the reasoning behind this approach is that in *vivo* a number of amino acid sites most probably undergo positive selection.

8) Any gene that shows evidence for being under the influence of positive selection is checked to ensure that results are not the result of misalignment or recombination. This step is performed using the alignment editor SE-AL (http://evolve.zoo.ox.ac.uk/software.html?id=seal) and PLATO 2.0 (Grassly and Holmes 1997) respectively. This step is not automated and needs input from the user. PLATO locates spatial variation in an alignment that results in different regions of the alignment supporting different phylogenies. Such variation is indicative of recombination. PLATO 2.0 utilises the maximum likelihood phylogenetic tree for a gene together with its substitution model and calculates the likelihood of this hypothesis along the alignment. MODELTEST 3.04 (Posada and Crandall 1998) was used to determine the optimum model of sequence evolution and associated parameters. Using PAUP* 4.0b10 (Swofford 1998) these parameters were used to infer a tree using the ML framework. Regions of the alignment that have the lowest likelihood are then tested for significant departure from the null hypothesis using Monte Carlo simulation; significance indicates the failure of the null model to explain the observed data (Grassly and Holmes 1997).
9) All possible pairwise distances (6 in the case of 4 taxa) were calculated for the codon based nucleotide alignments from step 4 above. The distances methods of Yang and Goldman (Goldman and Yang 1994) and Nei and Gojobori (Nei and Gojobori 1986) were used in this step of the analysis.
10) If dN was larger than dS for more than half of the sequence pairs compared in a particular gene group then this group was regarded as a candidate gene on which positive selection operates, this is the same criteria used by Endo *et al* (1996).

2.2.3 Simulation studies

The power and accuracy of LRTs have been investigated in the past (Anisimova *et al.* 2001). Some general observations from this study were that for the LRT to be powerful a minimum of six sequences should be used in any analysis. Sequence length and divergence will also have an impact on the analyis. All single gene families in this analysis contained only four sequences; obviously, this raises concerns regarding type 1 (false positive) errors. These concerns were examined by simulating replicate data sets under the null hypothesis of neutral evolution and analyzing them using both the null and the alternative positive selection hypothesis. The distribution of the test statistic $2\Delta l$ among replicates is compared to a χ^2 distribution with 2 free parameters for each of the three comparisons (i.e. M0 *vs* M3, M1 *vs* M2, M7 *vs* M8). One hundred pseudo data sets were simulated for families containing 4 taxa and varying sequence length sequence (200, 400 and 1000 codon positions) using the program EVOLVER from the PAML software package (Yang 1997). This program generates a codon sequence at the root of an inferred tree and evolves sequences along the branches of the phylogeny using specified branch lengths and substitution parameters (Yang 1997). The number of times that the ML method found a pseudo dataset to be evolving under the influence of positive selection represents how often type 1 errors may occur. Using the positive selection pipeline described above I again tested for evidence of positive Darwinian selection in these pseudo datasets.

To ensure that there are no systematic biases in the genes that show evidence for positive selection, a second simulation study was performed. From the ML *Neisseria* analysis, the 75 genes that were shown to be evolving under positive selection (see results) were selected. One hundred pseudo sequences of equal length were generated for each of the 75 gene families using EVOLVER along the appropriate ML tree, the observed codon frequencies (found using the program GCUA 1.1 (McInerney 1998)) and parameters such as transition:transversion ratio (reported in the initial ML analysis) were also determined and incorporated . This procedure was carried out on all 75 genes. The positive selection pipeline described above was used again to determine the frequency of false positives.

2.3 Results

2.3.1 *Neisseria* dataset

N. meningitidis serogroup B was used as a query genome against the *Neisseria* database which contained 10,002 ORFs. From the 2,156 BLAST searches that were performed 1,190 single gene families where all four taxa are present were located. The average length of each gene family after alignment was found to be 338 codon positions, with the largest family having 896 aligned codon positions while the shortest was 100 aligned codon positions. General functional categories of each gene family were determined using the nomencluture of The Institute for Genomic Research comprehensive microbial database (http://www.tigr.org). Functions for all genes in all datasets reported herein were determined in this manner. From the 1,190 single gene families it is noticeable that a large proportion of them are hypothetical proteins (26.3%) (Table 2.1). From the ML analysis of positive selection, strict criteria were set in order for a gene family to be a candidate for being under the influence of positive selection (see section 2.2.2). Of the 1,190 single gene families 75 (6.3%) (Table 2.1) were found to meet the strict criteria set above. No single class of protein function was shown to be atypically under the influence of positive selection as such genes were distributed across all function types (Table 2.1). If all genes that have a number of sites with $\omega > 1.3$ are considered (i.e. removing criteria that 5% of sites must have $\omega >1.3$) then it is found that 18.57% of genes exhibit evidence of positive Darwinian selection in at least one codon (Table 2.1). From the pairwise comparisons and using the same criteria as Endo *et al* (1996) only a single *Neisseria* gene was inferred to be under the influence of positive selection. This gene (NMB0332) is associated with the cell envelope and was also detected by the ML analysis. NMB0332 is involved in pilin production, as is NMB0887 which was also detected by the ML analysis as undergoing positive selection. Pili are filamentous organelles that bind to host cells. A typical bacterium has approximately 500 of these filaments projecting from the membrane surface. As the pili are membrane exposed it follows that they must be under intense scrutiny by the host immune system (Perry *et al.* 1988). It makes biological sense

Table 2.1: List of broad functions and frequency of occurrence for the 1190 *Neisseria* single gene families. The third column relates to the number of genes that have at least one amino acid site with ω > 1.3. The last column refers to the number of genes that exhibit evidence of positive selection using the strict criteria set out in the methods. In this scenario at least 5% of sites must have ω > 1.3.

Function	Total	ω > 1.3[a]	ω > 1.3[b]
Amino acid	60	14(23%)	5 (8%)
Biosynthesis of Cofactors	63	8 (13%)	3 (5%)
Cell processes	51	6 (12%)	1 (2%)
Cell envelope	67	13 (19%)	7 (10%)
CI metabolism	30	7 (23%)	3 (10%)
DNA metabolism	67	13 (19%)	6 (9%)
Energy metabolism	122	17 (14%)	2 (2%)
Fatty acid metabolism	19	2 (11%)	0 (0%)
Hypothetical	313	65 (21%)	26 (8%)
Other	6	3 (50%)	2 (33%)
Protein Fate	49	8 (16%)	3 (6%)
Protein synthesis	79	12 (12%)	2 (2%)
Purines etc	38	7 (18%)	1 (3%)
Regulatory	39	7 (18%)	3 (8%)
Transcription	31	3 (10%)	0 (0%)
Transport & Binding	70	19 (22%)	5 (6%)
Unclassified	1	0 (0%)	0 (0%)
Unknown	85	17 (20%)	6 (7%)
Totals	**1190**	**221 (18.5%)**	**75 (6.3%)**

for the pili to have undergone positive selection in an effort to evade host antibodies that may be directed against them. Evidence of strong positive selection for the pilE gene of *N. meningitidis* pili has been demonstrated in the past (Andrews and Gojobori 2004). The pairwise method did not infer positive selection in any of the other four datasets and therefore will not be mentioned until later discussions.

From the 75 *Neisseria* genes that exhibit evidence of having undergone positive selection according to the ML method we wished to ascertain in how many genes does the event that caused positive selection appear to be the result of speciation? The *Neisseria* dataset contains three commensals of the nasopharynx (*N. meningitidis* serogroups A, B and C) and also a commensal of the urogenitary tract (*N. gonorrohoeae*). Perhaps positive selection of particular metabolic pathways have enabled the *Neisseria* to diversify into alternative niches. To test this hypothesis we utilised the branch site models of Yang (Yang and Swanson 2002). These models are extensions of two of the site-specific models and can account for heterogeneous selective pressures at different sites and among lineages. According to these models, 40 (53%) of the 75 genes may be the result of the speciation event between the *N. meningitidis* strains and *N. gonorrohoeae* as the branch site models described the data better than the competing site-specific models (Table 2.2). The next step in the analysis was to determine if there were any functional relationships between the 40 genes inferred by the branch site models. A search of the Kyoto encyclopaedia genome database (Kanehisa *et al.* 2004) yielded information regarding the functional pathways in which 11 of the genes are involved (Table 2.2). The remaining 29 families were either hypothetical proteins or the pathway in which they are involved have yet to be described. Of the 11 genes which pathway information is available, two are involved in the biosynthesis of tryptophan (Figure 2.2).

Recent criticisms have found the above ML site branch site models to be too liberal in detecting positive selection. As a precaution a parsimony method that can test for evidence of positive selection as implemented in the software CRANN 1.04 (Creevey and McInerney 2003) was used. This method can detect positive selection along lineages and

Table 2.2: *Neisseria* genes that exhibit evidence of being under the influence of positive selection according to the ML branch site models. These genes may be involved in the speciation event that distinguishes *N. meningitidis* from N. *gonorrohoeae* 15 hypothetical genes are not shown. The pathways these genes are involved in are listed, the biological pathways for all genes of interest have not yet been determined. Underlined genes are those that have been corroborated by parsimony method and are result of peciation between *N. gonnorrohoeae* and other *Neisseria* strains.

Category	Gene Identifier and function	Pathway
Amino acid	NMB0627 phosphoribosyl-AMP cyclohydrolase	nme0340 Histidine metabolism
Amino acid	NMB0688 N-(5'-phosphoribosyl)anthranilate isomerase	nme0400 Tryptophan biosynthesis
Amino acid	NMB1446 3-dehydroquinate dehydratase	nme0400 Tryptophan biosynthesis
Cofactors biosynthesis	NMB0777 uroporphyrinogen-III synthase	nme0860 Porphyrin metabolism
Cofactors biosynthesis	NMB1616 phosphomethylpyrimidine kinase	nme0730 Thiamine metabolism
Cell Envelope	NMB0285	N/A
Cell Envelope	NMB0458 glutamate racemase	nme0471 D-Glutamate metabolism
Cell Envelope	NMB0890 type IV pilin-related protein	nme3090 Type II excretion system
Cell Envelope	NMB2032	N/A
Dna metabolism	NMB0076	N/A
Dna metabolism	NMB0835 type I restriction enzyme protein,	N/A
Dna metabolism	NMB1375	N/A
Energy	NMB0717 cytochrome	N/A
Protein fate	NMB0622 outer membrane lipoprotein carrier protein	N/A
Protein fate	NMB0791 peptidyl-prolyl cis-trans isomerase	N/A
Protein fate	NMB1832 lipoprotein signal peptidase (lspA)	nme3060 Protein transport
Protein synthesis	NMB0721 translation initiation factor 3 (infC)	N/A
Purines	NMB1996 phosphoribosylformylglycinamidine synthase	nme0230 Purine metabolism
Regulator	NMB0748 host factor-I (hfq)	N/A
Transport protein	NMB0881 sulfate ABC transporter, permease protein	nme2010 ABC transport system
Transport protein	NMB1783	N/A
Unknown	NMB0643 MafB-related protein	N/A
Unknown	NMB0833 type I restriction protein	N/A
Unknown	NMB1448 DNA-damage-inducible protein	N/A
Unknown	NMB1952 stringent starvation protein B (sspB)	N/A

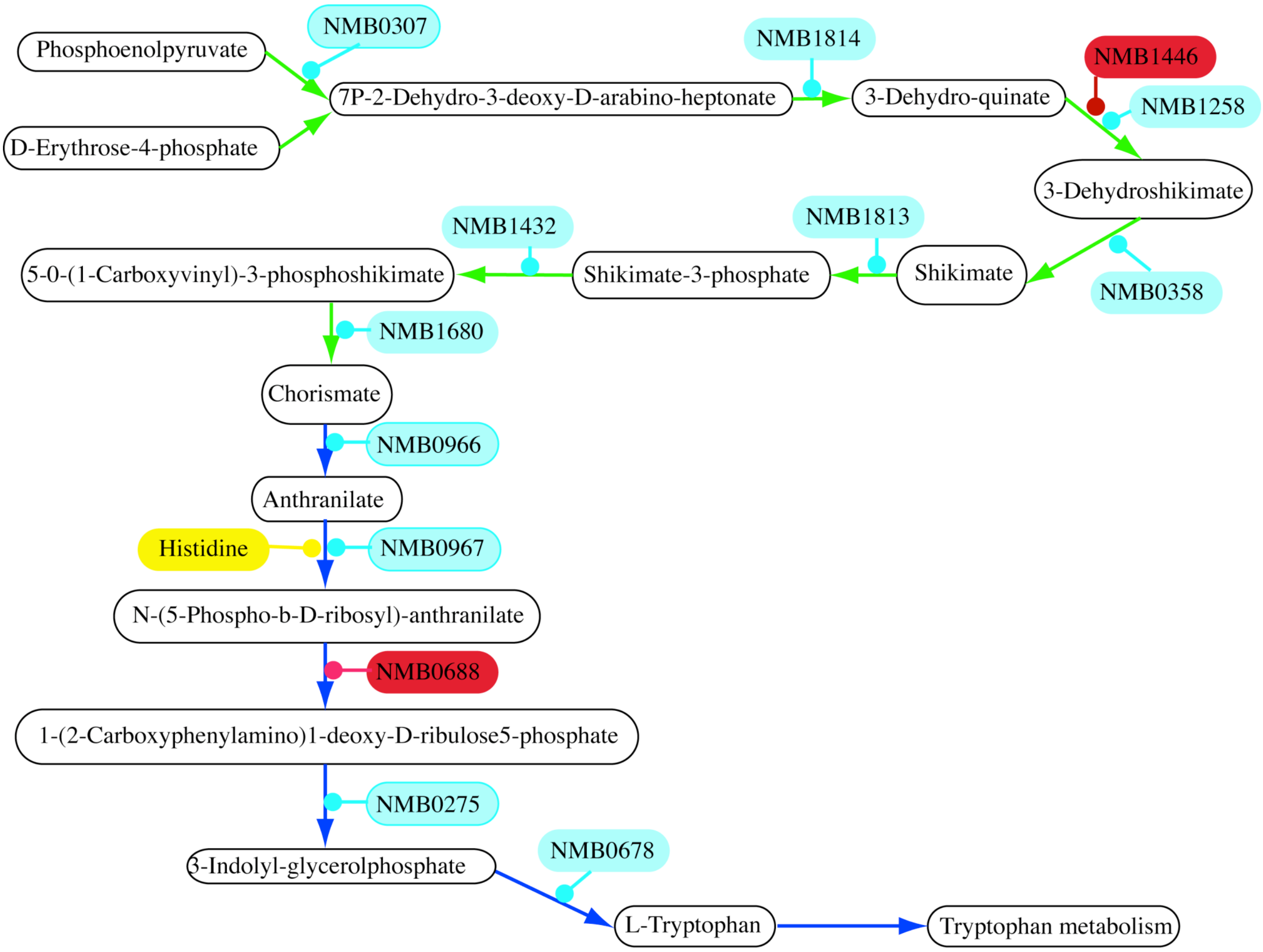


Figure 2.2: Simplified diagram of the Kyoto Encyclopaedia of Genes and Genomes pathway nme0400, which describes the biosynthesis of tryptophan and also the shikimate pathway. Two families that exhibit positive selection along the *N. gonorrohoeae* lineage are involved in this pathway and are highlighted in red. These are NMB1446 (3-dehydroquinate dehydratase I) and NMB0688 (N-5-phosphoribosyl-anthranilate isomerase) respectively.

operates by identifying all substitutions along a phylogeny and determining if they are synonymous or non-synonymous substitutions. Hypothetical ancestors are reconstructed at each internal node and the number of changes in descendent clades are counted. This results in four kinds of substitution types that are classified as either synonymous/non-synonymous invariable and synonymous/non-synonymous variable. Substitutions are considered invariable if the new character state is preserved in all subsequent lineages. Using a G-test the ratio of synonymous variable to synonymous invariable substitutions is compared to their nonsynonymous counterparts with the expectation that the ratio should not be significantly different. In the event of a significant difference between the two ratios it is possible to determine if there is an excess of nonsynonymous invariable substitutions a phenomenon indicative of positive selection. From the 40 genes inferred to be under the influence of positive selection by the ML method, 22 were also inferred by the parsimony method, for all genes *N.meningitidis* C was chosen as an arbituary outgroup. The genes involved in tryptophan biosynthesis were inferred by both methods.

2.3.2 Bacterial datasets

E. coli (06:k2:H1) was used as the query genome for BLAST searches against the *E. coli* database, which contained 19,370 ORFs. From the 5,533 BLAST searches performed, 1,734 single gene families where all four taxa are present were identified. The average length of each gene family after alignment was found to be 290 codon positions, with the largest family having 1,653 aligned codon positions while the shortest was 44 aligned codon positions. Of the 1,734 single gene families 20 (~1.1%%) (Table 2.3) were found to meet the criteria set out for positive selection. No single class of protein function was shown to have an unusually high proportion of genes under the influence of positive selection as such genes were distributed across all functional types (Table 2.3). Two genes involved in the cell envelope were found to be under the influence of adaptive evolution. The first was Ecs4499 which is involved in the biosynthesis of lipopolysaccharide while the second Ecs3217 is a pilin protein. As already mentioned the pili of bacteria protrude from the cell surface and are in direct contact with host immune system. Therefore conformational change through nonsynonymous substitutions should

Table 2.3: List of broad functions and frequency of occurrence for the 1732 *E. coli* single gene families. The third column relates to the number of genes that have at least one amino acid site with ω > 1.3. The last column refers to the number of genes that exhibit evidence of positive selection using the strict criteria set out in the methods. In this scenario at least 5% of sites must have ω > 1.3.

Function	Total	$\omega > 1.3^{a}$	$\omega > 1.3^{b}$
Amino acid	54	0(0%)	0(0%)
Biosynthesis of Cofactors	83	4(5%)	1(1%)
Cell processes	110	2(2%)	1(1%)
CI metabolism	43	1(2%)	1(2%)
DNA metabolism	72	3(4%)	3(4%)
Energy metabolism	163	1(1%)	1(1%)
Fatty acid metabolism	28	1(4%)	1(4%)
Hypothetical	588	23(4%)	7(1%)
Other	7	0(0%)	0(0%)
Protein Fate	55	1(2%)	0(0%)
Protein synthesis	87	0(0%)	0(0%)
Cell envelope	85	2(2%)	2(2%)
Purines etc	46	0(0%)	0(0%)
Regulatory	66	4(6%)	2(3%)
Transcription	20	1(5%)	0(0%)
Transport & Binding	77	3(4%)	0(0%)
Unclassified	77	1(1%)	0(0%)
Unknown	70	1(1%)	1(1%)
Viral Functions	1	0(0%)	0(0%)
Totals	**1732**	**48 (2.77%)**	**20 (1.15%)**

lead to increased fitness as the bacterium will be able to evade host antibodies directed against it.

The *Bacillus* dataset consisted of 19,871 ORFs and using *B. cereus* as the query genome resulted in 5,884 BLAST database searches, 759 single gene families were subsequently identified and of these only 18 (~2.4%) (Table 2.4) exhibited evidence of being under the influence of positive Darwinian selection. The average length of each gene family after alignment was found to be 262 codon positions, with the largest family having 1,232 aligned codon positions while the shortest was 56 aligned codon positions. Approximately 10% of some function types such as fatty acid metabolism and transcription exhibit evidence of adaptive evolution. These percentages are inflated however as only a small number of genes fit into these functions (13 and 19 respectively). Closer examination of the single transcriptional gene (Bsu3060) shown to be evolving under positive selection revealed that it is involved in Quorum sensing. Quorum sensing is a process of bacterial cell-to-cell communication involving the production and detection of extracellular signalling molecules called autoinducers. Quorum sensing has been shown to control competence in *Bacillus subtilis* (Perego and Hoch 1996). Bsu3060 produces a protein (*LuxS*) that regulates the production of autoinducer-2 (AI-2). AI-2 controls an assortment of niche specific genes such as motility in *E. coli* (Giron *et al.* 2002) and bacteremic infection in *N. meningitidis* (Winzer *et al.* 2002). So why is Bsu3060 under the influence of positive selection? A possible reason could include better regulation of AI-2. This in turn could increase the bacterium's ability to detect other bacteria in its vicinity through increased quorum sensing; therefore collectively controlling gene expression and synchronizing group behaviour which should increase group fitness (Xavier and Bassler 2003).

The *Chlamydia* dataset consisted of 4,255 ORFs and 643 single gene families of which 13 (2%) (Table 2.5) exhibited evidence of having undergone an adaptive event. As with all the datasets analysed in this study, no single category of gene seemed to undergo positive selection at a significantly higher rate than the other categories. *C. pneumoniae* (AR3909) was used as the query genome for database searches in this dataset. The

Table 2.4: List of broad functions and frequency of occurrence for the 759 *Bacillus* single gene families. The third column relates to the number of genes that have at least one amino acid site with ω > 1.3. The last column refers to the number of genes that exhibit evidence of positive selection using the strict criteria set out in the methods. In this scenario at least 5% of sites must have ω > 1.3.

Function	Total	ω > 1.3[a]	ω > 1.3[b]
Amino acid	42	0(0%)	0(0%)
Biosynthesis of Cofactors	37	1(3%)	1(3%)
Cell processes	59	2(3%)	2(3%)
CI metabolism	11	0(0%)	0(0%)
DNA metabolism	36	2(6%)	1(3%)
Energy metabolism	55	4(7%)	4(7%)
Fatty acid metabolism	13	1(8%)	1(8%)
Hypothetical	54	1(2%)	1(2%)
Other	4	0(0%)	0(0%)
Protein Fate	27	0(0%)	0(0%)
Protein synthesis	66	1(2%)	0(0%)
Cell envelope	26	0(0%)	0(0%)
Purines etc	17	0(0%)	0(0%)
Regulatory	32	0(0%)	0(0%)
Transcription	19	2(10%)	1(10%)
Transport & Binding	32	1(3%)	1(3%)
Unclassified	204	5(2%)	5(2%)
Unknown	21	1(5%)	1(5%)
Total	**759**	**21(3%)**	**18(2%)**

Table 2.5: List of broad functions and frequency of occurrence for the 643 *Chlamydia* single gene families. The third column relates to the number of genes that have at least one amino acid site with ω > 1.3. The last column refers to the number of genes that exhibit evidence of positive selection using the strict criteria set out in the methods. In this scenario at least 5% of sites must have ω > 1.3.

Function	Total	ω > 1.3[a]	ω > 1.3[b]
Amino acid	9	0(0%)	0(0%)
Biosynthesis of Cofactors	27	0(0%)	0(0%)
Cell processes	31	2(6%)	1(3%)
CI metabolism	9	1(11%)	0(0%)
DNA metabolism	35	3(9%)	0(6%)
Energy metabolism	38	2(5%)	1(3%)
Fatty acid metabolism	13	0(0%)	0(0%)
Hypothetical	257	13(5%)	8(3%)
Protein Fate	31	2(6%)	2(6%)
Protein synthesis	76	4(5%)	0(1%)
Cell envelope	29	2(7%)	0(3%)
Purines etc1	2	0(0%)	0(0%)
Regulatory	9	1(11%)	1(11%)
Transcription	12	0(0%)	0(0%)
Transport & Binding	34	0(0%)	0(0%)
Unknown	20	0(0%)	0(0%)
Total	**643**	**30(5%)**	**13(2%)**

average length of each gene family after alignment was found to be 340 codon positions, with the largest family having 1,570 aligned codon positions while the shortest was 52 aligned codon positions. A single gene CP1054 (GspF) is classified as being involved in protein fate (Table 2.5). Closer examination reveals that this gene is involved in type II secretion. Type II secretion allows bacteria to deliver virulence into their surroundings, it involves at least 12 genes and is required for translocation of secreted proteins across the outer membrane (Py *et al.* 2001). GspF is associated with both inner and outer membranes, as with the *Neisseria* and *E. coli* pilin proteins described above this means that GspF is possibly exposed to host immune responses. Positive selection of a number of functionally important sites on this protein may be essential so that the bacterium can evade host antibodies by altering possible antigen binding sites.

2.3.3 Neutral simulation studies

The aim of the first simulation study performed was to investigate the occurrence of type1 errors when only four taxa were present in a given dataset. The effect of sequence length was also examined. This was achieved by generating 100 pseudo datasets for three hypothetical cases. All three cases were similar in that they contained 4 taxa but the sequence length differed in each case (i.e. 200, 400 and 1,000 codon positions). The Pseudo datasets were artificially evolved in a manner that should not exhibit evidence of positive Darwinian selection. The occurrence of type 2 errors could not be investigated, as presently there is no available software to simulate adaptive evolution. The ML analysis was found to be quite robust in this simulation. For the pseudo datasets that are 200 codons in length only 3 of the100 pseudo datasets were found to exhibit evidence of positive selection according to the strict criteria specified earlier. With multiple testing one would expect to observe a small number of false positives being reported. This seems to be the case in this study. However the number of false positives reported is statistically insignificant according to a 95% confidence interval. When the sequence length was increased (i.e. to 400 and 1,000 codons respectively), no false positives were reported.

The aim of the second simulation performed was to ensure that possible systematic biases found within the real genes of this study were not affecting the ML analysis. To achieve this, sequences were evolved along the appropriate ML tree for each of the 75 *Neisseria* genes that were candidates for having undergone positive selection. Parameters such as codon frequency, transition:transversion ratio along with sequence length were used by EVOLVER to create pseudo datasets that should contain any systematic biases found within the real families. As with the first simulation these sequences are evolved in a manner that should not exhibit evidence of positive selection. From this second simulation study an insignificant number of the pseudo datasets (i.e. 5 or less for each gene family) showed evidence of being under the influence of positive selection.

2.4 Discussion

Kimura's neutral theory of evolution predicts that the majority of polymorphisms between and within species are neutral with respect to fitness and are the result of stochastic fluctuations within a finite population (Kimura 1979). The neutral theory has had a significant impact on our preconceptions as to what happens when mutation occurs. Before Kimura put his theory forward it was accepted that the majority of variation between species and populations was the result of natural selection. The neutral theory changed the way molecular evolutionists viewed polymorphisms and has lessened the role of natural selection. Recently a number of tests e.g. (Messier and Stewart 1997; Suzuki and Gojobori 1999; Yang *et al.* 2000; Creevey and McInerney 2002; Fares *et al.* 2002) have been developed to detect positive selection at the molecular level. A growing list of genes evolving under the influence of positive selection have been described using these methods e.g. (Swanson *et al.* 2001; Urwin *et al.* 2002; Lynn *et al.* 2004). The majority of such genes are either involved in sexual reproduction or host-pathogen interactions. The reasons that have been put forward as to why such genes have evolved by positive selection are the implicit needs for genes to continually improve fitness and evade host immune detection. More recently, a number of these tests have been applied on a genome wide scale and the results show that 5.3% of chordate and 3.6% of embryophyta gene families show evidence of having undergone positive selection (Liberles *et al.* 2001). Other studies have estimated that 25% of all amino acid substitutions between *Drosophila simulans* and *Drosophila yakuba* have been fixed by natural selection (Bierne and Eyre-Walker 2004) while approximately 35% of all substitutions between primates have also been fixed by natural selection (Fay *et al.* 2001). Scientific theories go in cycles, before Kimuras neutral theory selectionist arguesments dominated molecular evolution theory (Liberles and Wayne 2002). It would appear that as more genome sequence data becomes available the tide seems to be shifting back towards a selectionist view (Liberles and Wayne 2002).

In order to enter this selectionist/neutralist debate an investigation into the prevalence of positive selection in bacterial populations using a ML approach was performed. From the single genes tested, the *Neisseria* dataset was shown to exhibit the largest proportion of genes (~6.3%) having evidence of being under the influence of positive Darwinian selection, while the *E. coli* dataset had the smallest proportion (~1.1%). Of the positively evolving genes from all the groups examined I expected to see a significant skew in the frequency of membrane associated genes as all datasets contained human pathogens and as such should be involved in the "evolutionary arms race" (Urwin *et al.* 2002) between host and pathogen. No single class of gene was shown to disproportionably exhibit evidence of positive Darwinian selection as those inferred were distributed evenly over different function types. From the 75 *Neisseria* genes found to be evolving under the influence of adaptive evolution I wished to determine the number that may be the result of a speciation event between the nasopharynx commensals and the urogenitary bacterium. This was achieved by using modified ML methods that can account for positively selected sites along particular lineages (*N. gonorrohoeae* in this case). Of the 75 genes that were shown to exhibit positive selection 40 are a result of differences between the nasopharynx bacteria and urogenitary bacterium according to the ML branch site models. The functions and pathways for all 40 of these genes have yet to be described. However, they include genes involved in the cell envelope and amino acid metabolism. Two of the genes implicated by both ML branch site models and parsimony models to be under the influence of positive Darwinian selection are involved in the biosynthesis of tryptophan (Figure 2.2). The first gene found involved in this pathway was 3-dehydroquinate dehydratase I (NMB1446). 3-dehydroquinate dehydratase I is involved in the conversion of 3-dehydroquinate to 3-dehydroshikimate. This gene is not directly involved in tryptophan production but plays an important role in the shikimate pathway. The shikimate pathway involves seven enzymatic reactions whose end product is chorismate, a precursor for aromatic amino acid biosynthesis. It has been suggested that chorismate is often a limiting factor in the formation of tryptophan. The second gene involved in the biosynthesis of tryptophan found to be under the influence of positive selection was N-5-phosphoribosyl-anthranilate isomerase (NMB0688). N-5-phosphoribosyl-anthranilate isomerase is involved in the formation of 1-2-carboxyphenylamino-1-deoxy-D-ribulose5-phosphate from N-5-phospho-β-D-ribosyl-anthranilate. Tryptophan is biochemically the most expensive of the amino acids to synthesize (Xie *et al.* 2002). Therefore selection of a number of genes involved in the biosynthesis of this amino acid may help increase the efficiency of this pathway thus providing a selective advantage when this amino acid is required. Another possibility is that *N. gonnorrohoeae* may be exposed to high levels of the T-cell derived pro-inflammatory cytokine IFN-γ. This cytokine induces a variety of biochemical changes in host metabolism apparently designed to thwart the ability of pathogens to gain access to host resources including L-tryptophan (Xie *et al.* 2002). It should be noted that there have been recent concerns regarding false positives being reported by the branch site models used in this analysis therefore, caution is needed when interpreting these results. When the 40 genes selected by the branch site models were reanalysed using a relative rate test (Creevey and McInerney 2002), 22 were shown to exhibit evidence of positive selection along the *N. gonorrohoeae* branch. The two genes involved in tryptophan biosynthesis are included in this set of 22 genes.

To address concerns regarding the small number of sequences in each data set (4 in each) and the affect it may have on the power of the likelihood ratio test a simulation of neutrality was performed. Datasets containing four sequences of varying lengths (200, 400 and 1000 codons) were created. The number of times that a dataset was reported to exhibit positive selection represented how often one would expect to have a type 1 error.

Using the same criteria applied to the real data only three false positives were found for the 200 codon dataset and none for the other two datasets. Similarly, datasets of equal length and compositional biases as the 75 *Neisseria* genes found to be under the influence of positive Darwinian selection were simulated, again an insignificant number of false positives were detected. It was hoped that these simulation studies would lend some support towards the results of the real datasets presented in this analysis. The results of the simulation studies demonstrate that when a dataset of four sequences, all of which are evolving neutrally are tested for evidence of positive selection the ML software used will rarely report type I errors.

In this chapter, the occurrence of positive selection in a number of bacterial groups is presented. To achieve this large scale search, an automated pipeline was created to perform all the relevant procedures, the main advantages of this pipeline are its usability (there is minimal input from the user) and repeatability. This pipeline could be used in future analysis for different bacterial groups and is not restricted to four taxa although an increase in taxon number will have computational costs especially with the ML analysis of positive selection. Overall, 4,324 single gene families from five genera were examined. According to the distance methods utilised, only one gene shows evidence of having undergone an adaptive event. This is not entirely unexpected as the distance methods that were used average the dN/dS ratio over the entire gene. Therefore, it is possible that amino acids that are evolving under positive selection will not be found if most of the gene is under purifying selection. Conversely, according to the ML analysis 126 (~3%) of these genes exhibit evidence of being under the influence of positive selection. The number of genes on which positive selection operates may be much greater than 3% as this analysis was restricted to four closely related strains in each dataset. However, based on the results to hand one can only conclude that positive selection is not as rampant as some authors suggest (Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004). Only 9 of the 126 (~7%) positively evolving genes are associated with the cell envelope according to the functional categories of the comprehensive microbial resource of The Institute of Genomic Research. This finding provides evidence that genes not necessarily associated or exposed to the bacterial outer membrane are also under pressure for change.

To conclude, the prevalence of positively selected genes within the datasets examined would not lead one to conclude that positive selection is the *modus operandi* for the bacterial genera examined. Additional sequence data and better methods of detection may alter this conclusion however.

Chapter 3 Evidence of positive Darwinian selection in Bacterial outer membrane proteins

3.1 Introduction

Patterns of nonsynonymous substitutions in the major histocompatibility complex of vertebrates are maintained by balancing selection (Hughes and Nei 1988; Hughes and Nei 1989). Balancing selection describes a selective regime where successive amino acid change makes a protein more efficient at performing a particular task (Hughes and Nei 1988; Creevey and McInerney 2002). It follows therefore that if the human immune system is in an "evolutionary arms race" with pathogens, then similar pressures will be evident among the immunoexposed proteins of pathogens. To date, accelerated rates of nonsynonymous nucleotide substitutions relative to synonymous nucleotide substitutions have been demonstrated for a number of membrane associated proteins (Smith *et al.* 1995; Fares *et al.* 2001; Jiggins *et al.* 2002; Urwin *et al.* 2002; Andrews and Gojobori 2004). Furthermore, from a large scale database search for genes on which positive selection operates, Endo *et al* (1996) found that nine of the 17 genes found to be under positive selection were surface proteins of parasites or viruses. In this chapter two separate analyses are presented. These analyses show that a number of *Neisseria* outer membrane proteins (OMP) and a highly conserved δ-proteobacterial OMP are under positive Darwinian selection for diversity, most likely in response to pressure from the human immune system.

3.1.1 δ-Proteobacteria Outer Membrane Protein 85 (OMP85)

The cell envelope of Gram-negative bacteria consists of an outer membrane (OM) and an inner membrane separated by the peptidoglycan containing periplasm, both membranes contain proteins (Voulhoux *et al.* 2003). Outer membrane proteins (OMPs) are synthesized in the cytoplasm by the Sec machinery (Manting and Driessen 2000), have a β-barrel structure consisting of an even number of 8 to 22 membrane spanning β-strands with an antiparallel topology, which are connected by long and short loops forming β–hairpin structures (Voulhoux and Tommassen 2004). After translocation, the mature protein is released into the periplasm where it folds and is subsequently inserted into the membrane. The molecular mechanism underlying the shuttle mechanism of transport of completed OMPs to OM has recently been explained. Initially it was thought OMPs were translocated through membrane adhesion zones known as Bayer's junctions (Bayer 1968). Today there is very strong evidence to suggest that there is a specific proteinaceous machinery, specifically OMP85 and soluble chaperones that can dock transiently with the membrane surface (Tamm *et al.* 2001; Kleinschmidt 2003).

OMP85 homologues have been shown to be present in all Gram-negative bacteria (Stephens and Lammel 2001; Genevrois *et al.* 2003; Gentle *et al.* 2004; Voulhoux and Tommassen 2004) and also eukaryote mitochondria where they are involved in mitochondrial biogenesis. OMP85 can be divided into two domains, a periplasmic NH2-terminal domain and a 12 stranded β-barrel domain at the COOH- terminus (Voulhoux *et al.* 2003). Knockout of OMP85 in *Neisseria* strains results in the accumulation of porin monomers, these OMPs are normally found as trimers (Voulhoux *et al.* 2003). OMP85 knockout also increases the accumulation of monomeric PilQ. This secretin which is involved in type IV pili formation normally forms stable high molecular weight oligomers (Wolfgang *et al.* 2000). Immunofluoresence microscopy on knockout *Neisseria* strains has illustrated the importance of OMP85. In the absence of this gene the binding of antibodies directed against a number of OMPs (PilQ, PorA and PorB) was considerably weaker than in wild types. Further still, electron-dense material accumulates in the periplasm of *Neisseria* OMP85 depleted strains and although the identity of this material could not be revealed (Genevrois *et al.* 2003) it is most likely misassembled OMPs (Voulhoux and Tommassen 2004). The fact that OMP85 is also located in the lipopolysaccharide biosynthetic operon (Genevrois *et al.* 2003; Voulhoux *et al.* 2003) reinforces the view that this gene has an important role in OMP biogenesis. As OMP85 is surface-accessible to anti-OMP85 antibodies it has been suggested that this protein should be investigated for its vaccine potential (Stephens and Lammel 2001; Genevrois *et* *al.* 2003; Voulhoux and Tommassen 2004). The results presented in this chapter located a number of regions that encode surface exposed loops that exhibit an accelerated rate of nonsynonymous substitution. This finding is compatible with positive Darwinian selection, while structurally important transmembrane regions appear to be under the influence of purifying selection.

3.1.2 *Neisseria* OMPs

Neisseria meningitidis is a gram-negative encapsulated β-proteobacterium that is naturally competent for transformation with DNA. It only thrives in the human host and is not known to colonize any other animal or environmental niches. Meningococcal carriage is very much more common than disease (Cartwright 1995). Asymptomatic colonization of the nasopharynx is common, reaching 10% or more of the population in many countries. However for reasons still poorly understood, certain strains can penetrate the mucosal epithelium and gain access to the circulation system (Grandi 2003). In individuals lacking humoral immunity to meningococci, proliferation of the organisms in the blood may lead to septicaemia, characterised by circulatory collapse, multiple organ failure and coagulopathy. Additionally some virulent strains may also reach the subarachnoid compartment and initiate meningitis (Nassif *et al.* 1999). Meningococcal meningitis is a significant public health problem and is responsible for deaths and disability through epidemics in sub-Saharan Africa, and many other sporadic cases worldwide (Tettelin *et al.* 2000). Without prompt antibiotic treatment meningococcal infection is almost always fatal (Nassif 2002) and even with prompt treatment death and sequelae are common (Naess *et al.* 1994). Those most at risk from meningococcal infection include persons in the 15-24 age group and children under the age of 5. Annual incidence of meningococcal disease varies from 0.5 to 10 per 100,000 persons (Schuchat *et al.* 1997) but during epidemics the incidence rate can rise to above 400 per 100,000 persons (Schuchat *et al.* 1997). Presently five pathogenic *N. meningitidis* serogroups (A, B, C, Y and W135) have been described based on capsular polysaccharide typing (Gotschlich *et al.* 1969).

To date there has been some success in vaccine design against four of the five pathogenic serogroups (A, C, Y and W135). These vaccines consist of purified polysaccharide antigens, are highly effective in adults and the high molecular polysaccharides used in these vaccines are produced in the same method as first described by Gotschlich (1969). There is presently no effective vaccine against serogroup B however, as the capsular polysaccharide of this bacterium is identical to a widely distributed human carbohydrate (polysialic acid) preventing the use of this as a vaccine antigen for fear it may induce autoimmunity (Nassif 2002). Vaccine design against strains of serogroup B is a priority as they cause the majority of invasive disease in Europe (>50% of cases) and the USA (>30% of cases).

The problems associated with developing a meningococcal serogroup B vaccine has led researchers to explore the possibility of targeting membrane exposed proteins. Administration of an immunizing agent early in life should induce the production of serum bactericidal antibodies, which are the mark of resistance of meningococcal disease (Goldschneider *et al.* 1969). Suitable antigens include those that are conserved among a population, are expressed on the surface of the bacteria and also induce the production of bactericidal antibodies (Nassif 2002). Using these criteria, the complete genome sequence of *N. meningitidis B* (strain MC58) (Tettelin *et al.* 2000) was used to locate open reading frames (ORFs) that potentially encode surface-exposed or exported proteins (Pizza *et al.* 2000). From the 570 candidate ORFs located, they successfully cloned and expressed 350 of these into *Escherichia coli*. The resultant recombinant proteins were used to immunize mice and the immune sera were tested for bactericidal activity, as this correlates with protection in humans (Pizza *et al.* 2000). Finally, from 85 proteins that elicit bactericidal activity the suitability of these proteins was tested as candidate antigens for conferring protection against heterologous meningococcal strains. The tests were carried out on 34 different *N. meningitidis* clinical strains isolated worldwide and over many years. This process resulted in the identification of seven proteins that confer protection against both homologous and heterologous meningococcal strains. The sera against these antigens is capable of killing all the meningococcal strains so far utilised in the complement-mediated bactericidal assay, thus making the antigens particularly promising for vaccine formulations. Presently phase 1 clinical studies are in progress to establish the ability of these antigens to induce bactericidal antibodies in humans (Grandi 2003). These genes are highly conserved which is an unusual finding considering that in three decades of studies, with one exception (Martin *et al.* 1997), only antigenically variable surface exposed proteins had been described (Pizza *et al.* 2000). Furthermore the frequency of recombination of these genes is relatively low (0.11 using the homoplasy test (Maynard-Smith and Smith 1998)), a level of recombination that is similar to that of *Neisseria* house keeping genes (Maynard-Smith and Smith 1998). From mathematical modelling of viral infection dynamics it has been suggested that conserved epitopes are more appropiate as vaccine targets than variable epitopes (Nowak *et al.* 1991). Therefore, it would be helpful if we could predict candidate epitopes computationally as it should accelerate the entire vaccine process. Observations suggest that the success and failure of present day vaccines may be linked to the presence/absence of positive selection. For example, annual vaccines against influenza A virus have to be developed due to its rapid antigenic change; conversely, the extremely successful poliovirus vaccine has been shown to be under strong negative selection. These observations would lead us to conclude that candidate vaccine targets should contain no positively selected sites a view that has been expressed by others (Suzuki 2004).

From a protein structure perspective, it is likely that there are two categories of proteins, those that are amenable to change and those that, for structural reasons cannot easily change. An ideal vaccine candidate should be in the second category and it should be possible to determine from analysing the evolutionary history of a set of sequences if change via positive selection is easy or difficult in those sequences. In this analysis I test the hypothesis that the seven gene families currently being considered as potential vaccine candidates (Pizza *et al.* 2000) show evidence of historical positive selection. The reason for proposing this situation is that these proteins are known to be expressed at high levels, are surface exposed and elicit a strong immune reaction during infection. I discuss my findings in light of the implications for vaccine design.

3.2 Materials and Methods

3.2.1 OMP85 Sequences

OMP85 homologues from 10 δ-proteobacteria were analysed. They were *Escherichia coli* gi_1786374, *Xanthomonas campestris* gi_21230822, *Xylella fastidiosa* gi_28198247, *Salmonella typhimurium* gi_16418729, *Pseudomonas aeruginosa* gi_9949808, *Pasteurella multocida* gi_12722432, *Vibrio cholerae* gi_9656813, *Haemophilus influenzae* gi_1573938, *Yersinia pestis* gi_15979119, *Shigella flexneri* gi_24050380. Homologues from other bacteria were available but the sample was restricted to the δ-proteobacteria in an effort to limit the effects of saturation of synonymous sites that occurs between distantly related sequences. Saturation of synonymous sites reduces the power of methods that detect positive Darwinian selection. The δ-proteobacteria were selected as they were the largest bacterial division in terms of available data.

3.2.2 *Neisseria* sequences

Twenty two *N. meningitidis* serogroup B strains, three serogroup A strains, two serogroup C strains, one strain each of serogroups W, X, Z and Y, one strain of *N. cinerea*, one strain of *N. lactamica* and three strains of *N. gonnorrhoeae* were used in this analysis (Table 3.1). Seven gene families (putative vaccine targets) were examined. The nomenclature of each family corresponds to the annotation of the completely sequenced MC58 genome (Tettelin *et al.* 2000) and therefore are called by the names NMB0033, NMB0992, NMB1162, NMB2002, and NMB2132. The accession numbers for each of the genes is as follows. NMB0033 (AF226387-AF226417 and AF235143-AF235145). NMB0992 (AF226356-AF226386). NMB1220 (AF226542-AF226572 and AF235157-AF235159). NMB1946 (AF226480-AF226510 and AF235151-AF235153). NMB2001 (AF226449-AF226479 and AF235148-AF235150). NMB2132 (AF226418-AF226448 and AF235146-AF235147). Any nucleotide sequence that was found to be identical to another strain was removed from the analysis.

Table 3.1: *Neisseria* strains used in this analysis. Serogrouping and virulence are shown in columns 4 and 5. Where information is available strain identification, country of origin and year of isolation are displayed in columns 1, 2 and 3 respectively.

Strain	Country	Year	Serogroup	Invasive/Noninvasive
A22	Norway	1986	W	Non-Invasive
297-0	Chile	1987	B	Non-Invasive
205900	Mali	2000	A	Invasive
Z2491	Gambia	unknown	A	Invasive
E32	Norway	1988	Z	Non-Invasive
BZ133	Holland	1982	B	Invasive
BZ147	Holland	1964	B	Invasive
BZ83	Holland	1984	B	Invasive
BZ232	Holland	1964	B	Invasive
BZ169	Holland	1985	B	Invasive
BZ198	Holland	1986	B	Invasive
NGH15	Norway	1987	B	Non-Invasive
NGH36	Norway	1988	B	Non-Invasive
NGH38	Norway	1988	B	Non-Invasive
NGP165	Norway	1975	B	Invasive
90/18311	Scotland	1990	C	Invasive
93/4286	England	1991	C	Invasive
NGE31	Norway	1988	B	Non-Invasive
NGE28	Norway	1988	B	Non-Invasive
E26	Norway	unknown	X	Non-Invasive
F6124	Chad	1990	A	Invasive
NG3/88	Norway	1988	B	Invasive
MC58	Scotland	unknown	B	Invasive
H44/76	Norway	1976	B	Invasive
NG6/88	Norway	1988	B	Invasive
NGF26	Norway	1989	B	Non-Invasive
860800	Holland	1986	Y	Invasive
528	USSR	1989	B	Invasive
2996	USSR	1988	B	Invasive
1000	USSR	1988	B	Invasive
SWZ107	Switzerland	1986	B	Invasive
NGF62	USA	1990	gonorrohoeae	Invasive
NG-SN4	unknown	unknown	gonorrohoeae	Invasive
FA1090	unknown	1988	gonorrohoeae	Invasive

3.2.3 Sequence Alignments

Alignments for both analyses were created using the following procedeure. Stop codons were removed from all sequences. The nucleotide sequences were then translated into their amino acid equivalents and aligned using the default settings of CLUSTALW ver1.82 (Thompson *et al.* 1994). Gaps created in the amino acid alignment were transposed back to the nucleotide sequences to gain a codon based alignment using the in-house JAVA program Putgaps (http://bioinf.may.ie/software/putgaps). Codon alignments were corrected for obvious alignment ambiguity using the alignment editor Se-Al 2.0a11 (http://evolve.zoo.ox.ac.uk/software.html?id=seal).

3.2.4 Data analysis

Phylogenies for all gene families were reconstructed using the maximum likelihood framework implemented in PAUP* 4.0b10 (Swofford 1998). Firstly the optimal model of sequence substitution was selected by comparing the likelihood scores using MODELTEST 3.04 (Posada and Crandall 1998). The optimum model of substitution for OMP85 was found to be TrNef+I+G. The optimal models of substitution for the vaccine candidates were found to be GTR+I+G for NMB0033, TrN+I for NMB0992, K80+I for NMB1162, GTR+I+G for NMB1220, TrN+I for NMB1946, TrN+I for NMB2001 and GTR+G for NMB2132. Phylogenetic hypotheses that utilise these substitution models and the estimated parameters from Modeltest were compared to trees that were reconstructed using a Bayesian framework as implemented in MR BAYES 3.0B4 (Huelsenbeck and Ronquist 2001). For the Bayesian analysis the markov chain was run for 500,000 generations and sampled every 100th generation. The first 500 samples were discarded as burn-in and the resultant trees were summarised using a majority rule consensus method implemented in PAUP* 4.0b10 (Swofford 1998). The proposed topologies from both methods were identical for all gene families.

In the absence of tertiary structure data the ConSeq (Berezin *et al.* 2004) web server was used for the identification of biologically important residues within all seven gene families. Functionally important residues are often solvent accessible and evolutionary conserved while structurally important residues are normally highly conserved and found within the protein core. Ideally vaccine targets should be from the latter category as they are accessible and not likely to change. Using ConSeq we hope to locate such exposed functionally important residues.

3.2.5 Detection of recombination

The method of Grassly and Holmes (1997) as implemented in PLATO 2.0 was used to search for blocks of sequences that have incongruent phylogentic topologies due to recombination. This precautionary step was taken as a recent study (Anisimova *et al.* 2003) has shown that high levels of recombination can increase type I errors in the heterogenous ML methods used in this analysis.

3.2.6 Analysis of selection

The maximum likelihood (ML) approach of Yang *et al* (2000) as implemented in the PAML package 3.13 (Yang 1997) was used to examine selection pressures acting on the *Neisseria* vaccine candidates as well as the OMP85 homologues. Models M0, M1, M2, M3, M7 & M8 were used. See section 2.2.2 for a detailed description of these methods.

Maximum likelihood models that allow for heterogeneity in the dN/dS ratio among lineages were also tested. The simplest model is the one ratio model M0. The most general model is the free ratio model which assumes as many ω parameters as the number of branches in the tree, this model is a parameter rich model (Yang 1998). Models that fit between these two extremes include the two-ratio model and three ratio models these allow predefined lineages to have a different ω value to the rest of the tree (Yang 1998). All of these methods were utilised in an effort to detect if positive selection has acted along specific lineages within the δ-proteobacteria OMP85 and vaccine candidate homologues.

Recent studies have shown that under certain conditions ML methods can be sensitive to violations of assumptions made in models that test for positive selection, these sensitivities under certain conditions can result in false positives (Suzuki and Nei 2001; Suzuki and Nei 2004). To account for this, a maximum-parsimony method that tests for adaptive evolution was applied. A sliding window procedure that uses the model of Li (Li 1993) as implemented in SWAPSC (Fares 2004) was used. This method infers a statistically optimum codon-window size and slides it along the alignment. Each window step is then tested for the significance of nonsynonymous substitutions to synonymous substitutions and the nonsynonymous to synonymous rate ratio ω; in this manner positively, neutrally or negatively evolving sites can be located. This method also has the ability to test for saturation of synonymous substitutions (Fares 2004), if any sites are highlighted they can be removed.

3.3 Results

3.3.1 Recombination and saturation analysis

The effects of recombination on methods that detect positive selection have been documented (Anisimova *et al.* 2003). In an effort to ensure that these pitfalls do not affect the results reported the method of Grassly and Holmes (1997) was utilised to ensure that no recombination had occurred within the dataset presented. This method analyzes an alignment for sequence blocks within the alignment that deviate significantly from a predefined topology (the ML tree found earlier). No blocks within the OMP85 homologues or the vaccine candidate genes were found to deviate significantly from the imposed phylogeny. Saturation of synonymous sites is also an important issue when trying to detect adaptive evolution events by the criterion that dN/dS is greater than 1, as saturation can lead to the underestimation of dS and an inflation of the dN/dS ratio (Lynn *et al.* 2004). The moving window program SWAPSC (Fares 2004) was used in an effort to test for saturation of synonymous sites. No saturated sites were located in the OMP85 homologues and approximately 97% of the sites show strong purifying selection. The vaccine candidate genes were also shown not to contain saturated synonymous sites and for all 7 gene families more than 95% of sites show strong evidence of strong purifying selection.

3.3.2 Analysis of selection on OMP85 homologues

3.3.2.1 Maximum likelihood analysis

Maximum likelihood analysis of the selective pressures acting on the OMP85 gene provided strong evidence for positive selection (Table 3.2). The one ratio model (M0) predicted a ω value of 0.0761, this would indicate very strong purifying selection, however LRTs indicate that models that allow for positively selected sites increased the likelihood scores when compared with models that don't permit positive selection (Table 3.2). M2 (selection) fits the data significantly better than M1 (neutral) and locates ~3% of all sites in a class that is under the influence of positive selection (ω = 6.1). M3 (discrete) and M8 (beta) also fit the data better than M0 (one ratio) and M7 (beta) respectively. Both of these models indicate that ~3% of sites to have a ω value of 3.65 and 4.98 respectively, a value indicative of strong positive selection. Bayesian posterior probabilities identified 14 positively selected codons for M2, M3 and M8, these codons were almost identical in all cases except M2 did not infer site 740. Codon based likelihood models that allow for different dN/dS ratios among evolutionary lineages (Yang 1998) were used so as to determine if particular lineages in the datasets are undergoing positive selection (Figure 3.1). In all cases, none of the LRTs performed allowed for positive selection on any one lineage exclusively (Table 3.3) therefore, there is no evidence of lineage specific positive selection within this dataset.

Table 3.2: Results of the ML analysis of OMP85 using a variety of models. p_1, p_2 and p_3 refer to the proportion of sites in categories 1, 2 and 3 respectively. ω refers to the *dN/dS* ratios in these categories of sites. p and q are beta estimates. Models M1 and M7 do not allow positively selected sites.

Model	Estimates of Parameters	ln L	2Δln L	Positively selected codons
M0 (one ratio)	$\omega_1 = 0.0761$	19967.76		None
M1 (neutral)	$p_1 = 0.15$ $p_2 = 0.85$	-20780.4		Not allowed
	$\omega_1 = 0$ $\omega_2 = 1$			
M2 (selection)	$p_1 = 0.15$ $p_2 = 0.81$ $p_3 = 0.03$	-19649.4	(M1 *vs* M2)	286, 564, 673, 674, 675, 676, 735,
	$\omega_1 = 0$ $\omega_2 = 1$ $\omega_3 = 6.1$		$p < 0.0005$	736, 737, 738, 740, 743, 745
M3 (discrete k=3)	$p_1 = 0.38$ $p_2 = 0.57$ p3 = 0.04	-19603.4	(M0 *vs* M3)	286, 564, 673, 674, 675, 676, 735,
	$\omega_1 = 0.02$ $\omega_2 = 0.1$ $\omega_3 =3.65$		$p < 0.0005$	736, 737, 738, 740, 742, 743, 745
M7 (beta)	p= 0.90064 q= 8.27904	-19632.1		Not allowed
M8 (beta & ω)	p = 1.08 q= 13.88	-19586.9	(M7 *vs* M8)	286, 564, 673, 674, 675, 676, 735,
	$p_2 = 0.038$ $\omega_2 = 4.98$		$p < 0.0005$	736, 737, 738, 740, 742, 743, 745

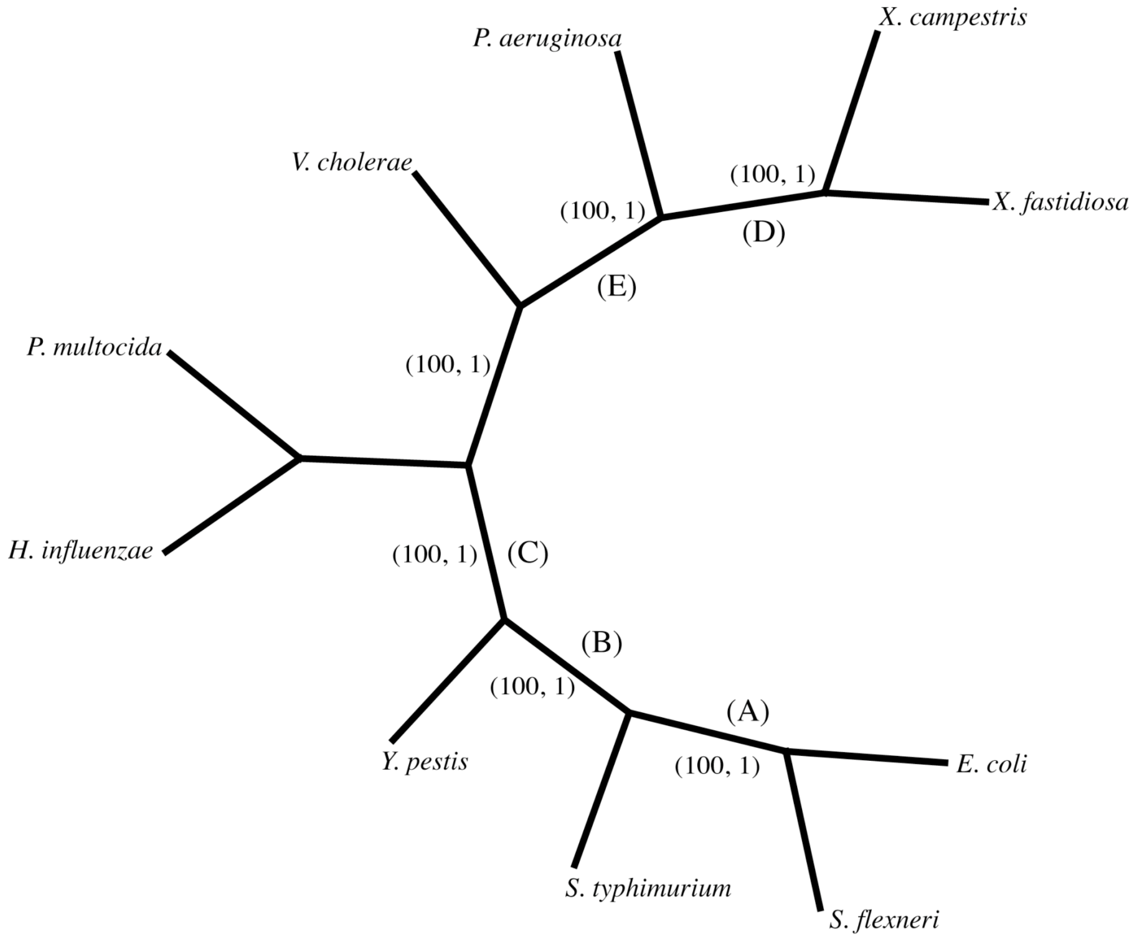


Figure 3.1: Maximum likelihood phylogenetic tree of OMP85 gene from 10 δ-protobacteria. Numbers in brackets correspond to 1000-pseudoreplicate bootstrap values and also branch supports from Bayesian analysis, both methods identified the same topology. Letters along branches correspond to the lineages tested as having a ω value that differs from the rest of the tree using the lineage specific methods of Yang (1998).

Table 3.3: Log likelihood values and parameter estimates using different branch models. The first column represents the model used and letters in brackets refer to the branch that was labelled (see Figure 3.1). The second column lists the number of parameters associated with each model. Log likelihood and transition:transversion ratio are listed in columns 3 and 4 respectively. ω_0 represents the background ω when a particular branch is allowed to vary from the rest of the tree while ω_1 corresponds to that variable ω. The one ratio model assumes that all branches in the tree have a uniform ω value, alternative models that allow for variable ω along predefined branches do not explain the data significantly better in any case and are therefore rejected. This leads to the conclusion that lineage specific positive selection is absent from this dataset. ω values for the freeratio model are not displayed as there are a great many of these

Model	*p*	lnL	*k*	ω_0	ω_1
One ratio	19	19967.76	1.422	0.0761	n/a
Two ratio (A)	20	19967.75	1.423	0.0721	0.032
Two ratio (B)	20	19967.76	1.430	0.0711	0.079
Two ratio (C)	20	19966.92	1.423	0.0721	0.069
Two ratio (D)	20	19967.60	1.433	0.0711	0.033
Two ratio (E)	20	19967.61	1.435	0.0760	0.063
Freeratio	45	19955.15	1.420	n/a	n/a

3.3.2.2 Parsimony analysis

All of the sites inferred to be under the influence of positive Darwinian selection using the ML method were also detected using the sliding window approach. The sliding window approach also inferred additional sites, which may be under the influence of positive Darwinian selection. The majority of these sites (15 in total) lie between codons 564-590. The ML method had also put these sites in a class with $\omega > 1$, but they were not considered significant as their Bayesian posterior probabilities were less than 0.95. From the positively selected amino acid sites an average ω value of 2.87 was found. Maximum parsimony methods are very conservative (Suzuki and Nei 2002) and may be subject to the problem of possible convergences (Lynn *et al.* 2004). Despite this problem, results from ML and MP methods show broad agreement with one another and therefore do not affect any conclusions derived from this study.

3.3.2.3 OMP85 secondary structure and location of positively selected sites

As all OMP85 homologues have been shown to be highly conserved (Voulhoux *et al.* 2003; Gentle *et al.* 2004; Voulhoux and Tommassen 2004) and in the absence of tertiary structures, the OMP85 topology of Voulhoux *et al* (2003) was used as a working model on which to plot the positively selected sites that were inferred by both methods. All but one positively selected site (site 281) are found in the regions that are membrane exposed, two loops in particular account for the majority (12 out of 14) of positively selected sites (Figure 3.2). The remaining positively selected site is located in the periplasmic NH_2-terminal domain.

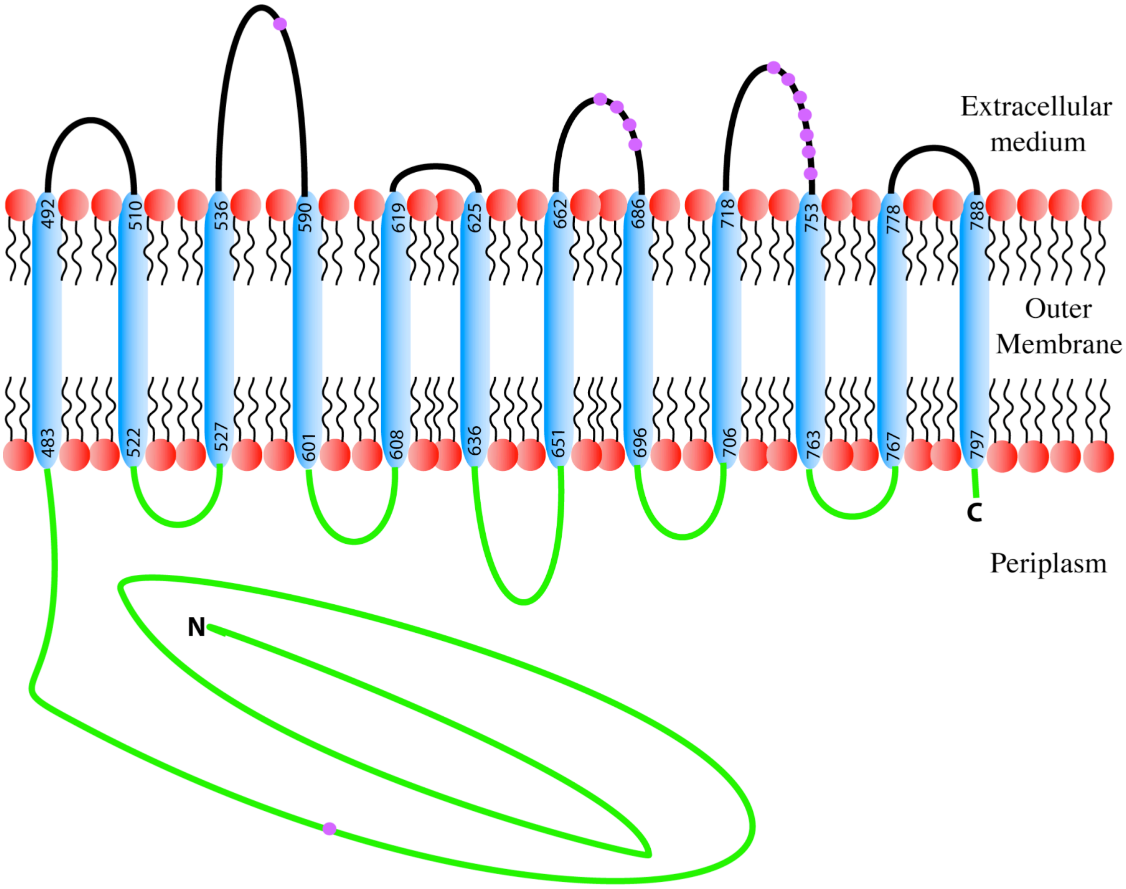


Figure 3.2: Topology of OMP85 as predicted by Voulhoux *et al* 2003. Periplasmic domains are green, exposed domains are black and β-sheets are blue. The first and last amino acid of each β-strand are indicated. Amino acid sites that are under the influence are shown as purple dots. The amino (N) and carboxy (C) termini of the protein are indicated. This model is not to scale.

3.3.3 Analysis of selection on vaccine candidate genes

3.3.3.1 Maximum likelihood analysis

Five of the seven vaccine candidate genes (NMB0033, NMB1162, NMB1220, NMB1946 and NMB2001) did not exhibit strong evidence of positive selection. In this analysis gene families were only placed into the category of undergoing positive selection if all 3 LRTs performed were significant. The estimated parameters and LRTs for these 5 gene families are shown in Tables 3.4, 3.5, 3.6, 3.7 & 3.8.

Analysis of the selective pressures acting on the remaining two families (NMB0992 and NMB2132) provided strong evidence for positive selection (Tables 3.9, 3.10). The three LRTs for the NMB0992 data indicated that there was strong evidence for positive selection at some sites. All three LRTs were highly significant (Table 3.9). M2 indicated that approximately 9% of sites had an ω value of ~3.95. For both M3 and M8 approximately 11% of sites fell into a strong positively selected class (ω = 3.67). The positively selected codons identified by Bayesian posterior probability were identical for M3 and M8. A smaller number were suggested by M2, but these were subsequently found to be a subset of those predicted by M3 and M8. The LRT analysis of site-specific variable selective pressures acting on NMB2132 again provided support for hypotheses of adaptive evolution (Table 3.10). M0 estimated an average ω value of 0.579. Again, using models that allow for variable ω values across sites, all three LRTs are highly significant ($p < 0.0005$) and all three models find more than 9% (9.6%, 30% and 10%, for models M2, M3 and M8 respectively) of sites to have an ω value greater than 1 (3.97, 2.13 and 3.66). Bayesian posterior probabilities indicate that the same 16 codons can be assigned with confidence ($p>0.95$) to this class of sites with a high ω value for M3 and M8 while a smaller number (13) are found for M2. These 13 sites overlap with those found by the other two models.

In order to test the hypothesis that one or more lineages are responsible for the finding of positive selection or to identify lineages where positive selection is stronger, likelihood

Table 3.4: ML analysis of NMB0033 using a variety of models. p_1, p_2 and p_3 refer to the proportion of sites in categories 1, 2 and 3 respectively. ω refers to the *dN*/*d*S ratios in these categories of sites. p and q are beta estimates. Models M1 and M7 do not allow positively selected sites.

Model	Estimates of parameters	ln L	2Δln L	Positively selected codons
M0 (one ratio)	$\omega_1 = 0.0791$	-3605.82		None
M1 (neutral)	$p_1 = 0.875$ $\omega_1 = 0.000$	-3563.69		Not allowed
	$p_2 = 0.124$ $\omega_2 = 1.000$			
M2 (selection)	$p_1 = 0.000$ $\omega_1 = 0.000$	-3553.12	(M1 *vs* M2) 21.14	None
	$p_2 = 0.068$ $\omega_2 = 1.000$		$p < 0.0005$	
	$p_3 = 0.931$ $\omega_3 = 0.021$			
M3 (discrete)	$p_1 = 0.939$ $\omega_1 = 0.024$	-3552.78	(M0 *vs* M3) 105.8	11, 48, 55, 112, 197, 207, 292, 312,
	$p_2 = 0.060$ $\omega_2 = 1.194$		$p < 0.0005$	338, 341, 342, 417, 421, 423
M7 (beta)	$p = 0.043$ $q = 0.40067$	-3555.92		Not allowed
M8 (beta + ω)	$p = 2.547$ $q= 99.00$	-3552.81	(M7 *vs* M8) 6.22	11, 48, 55, 112, 197, 207, 292, 312,
	$p_2= 0.05971$ $\omega_2=1.20$		$p < 0.05$	338, 341, 342, 417, 421, 423

Table 3.5: Results of the ML analysis of NMB1162 using a variety of models. p_1, p_2 and p_3 refer to the proportion of sites in categories 1, 2 and 3 respectively. ω refers to the *dN*/*d*S ratios in these categories of sites. p and q are beta estimates. Models M1 and M7 do not allow positively selected sites.

Model	Estimate of Parameters	ln L	2Δln L	Positively selected codons
M0 (one ratio)	ω_1=0.1680	-1195.00		None
M1 (neutral)	p_1=0.80 ω_1=0 p_2=0.20 ω_2=1	-1193.05	(M1 *vs* M2) 3.92	Not allowed
M2 (selection)	p_1=0.79 ω_1=0 p_2=0 ω_2=1 p_3=0.21 ω_3=0.903	-1193.04		None
M3 (discrete k=2)	p_1=0.79 ω_1=0 p_2=0.21 ω_2=0.90	-1193.04	(M0 *vs* M3) 3.92	None
M7 (beta)	p=0.013 q=0.063	-1193.04		Not allowed
M8 (beta + ω)	p= 0.00275 q= 2.06 p_2= 0.20 ω_1= 0.90	-1193.04	(M7 *vs* M8) 0	None

Table 3.6: Results of the ML analysis of NMB1220 using a variety of models. p_1, p_2 and p_3 refer to the proportion of sites in categories 1, 2 and 3 respectively. ω refers to the *dN*/*d*S ratios in these categories of sites. p and q are beta estimates. Models M1 and M7 do not allow positively selected sites.

Model	Estimate of Parameters	ln L	2Δln L	Positively selected codons
M0 (one ratio)	ω_1=0.0218	-1996.12		None
M1 (neutral)	p_1=0.975 ω_1=0 p_2=0.024 ω_2=1	-1986.03	(M1 *vs* M2) 0	Not allowed
M2 (selection)	p_1=0.98 ω_1=0 p_2=0.02 ω_2=1 p_3=0 ω_3=0	-1986.03		None
M3 (discrete k=2)	p_1=0.97711 ω_1=0 p_2=0.02289 ω_2=1.20	-1985.98	(M0 *vs* M3) 20.28 $p<0.0005$	9, 12, 137, 140, 170
M7 (beta)	p=0.010 q=0.24	-1989.14		Not allowed
M8 (beta + ω)	p=0.001 q=2.16 p_1=0.022 ω_1=1.18	-1985.98	(M7 *vs* M8) 6.32 $p<0.05$	9, 12, 137, 140, 170

Table 3.7: Results of the ML analysis of NMB1946 using a variety of models. p_1, p_2 and p_3 refer to the proportion of sites in categories 1, 2 and 3 respectively. ω refers to the *dN*/*d*S ratios in these categories of sites. p and q are beta estimates. Models M1 and M7 do not allow positively selected sites.

Model	Estimate of Parameters	ln L	2Δln L	Positively selected codons
M0 (one ratio)	ω_1=0.0817	-1485.26		None
M1 (neutral)	p_1=0.93 ω_1=0 p_2=0.07 ω_2=1	-1469.12	(M1 *vs* M2) 2.36	Not allowed
M2 (selection)	p_1=0.94 ω_1=0 p_2=0.044 ω_2=1 p_3=0.016 ω_3=3.73	-1467.94		None
M3 (discrete k=2)	p_1= 0.96 ω_1=0.013 p_2=0.04 ω_2=2.68	-1467.84	(M0 *vs* M3) 34.84 p<0.0005	42, 62,65, 75, 158
M7 (beta)	p=0.0026 q=0.027	-1469.83		
M8 (beta + ω)	p=0.001 q=2.127 p_1= 0.05 ω_1=1.78	-1468.16	(M7 *vs* M8) 3.34	33 , 36, 42, 62, 65, 75, 119, 158, 163, 278

Table 3.8: Results of the ML analysis of NMB2001 using a variety of models. p_1, p_2 and p_3 refer to the proportion of sites in categories 1, 2 and 3 respectively. ω refers to the *dN*/*d*S ratios in these categories of sites. p and q are beta estimates. Models M1 and M7 do not allow positively selected sites.

Model	Estimate of Parameters	ln L	2Δln L	Positively selected codons
M0 (one ratio)	ω_1=0.2443	-1182.66		None
M1 (neutral)	p_1= 0.80 ω_1=0 p_2= 0.20 ω_2=1	-1179.14	(M1 *vs* M2) 1.08	Not allowed
M2 (selection)	p_1=0.88 ω_1=0 p_2=0 ω_2=1 p_3=0.12 ω_3=2.00	-1178.60		21, 51, 80 , 83, 97, 126, 142, 152, 176, 241
M3 (discrete k=2)	p_1= 0.88 ω_1=0 p_2=0.12 ω_2=2.00	-1178.60	(M0 *vs* M3) 8.12 $p<0.05$	21, 51, 80, 83, 97, 126, 142, 152, 176, 241
M7 (beta)	p=0.001 q=0.0032	-1179.14		
M8 (beta + ω)	p=0.001 q=1.4759 p_1= 0.122 ω_1=2.00	-1178.60	(M7 *vs* M8) 1.08	21, 51, 80, 83, 97, 126, 142, 152, 176, 241

Table 3.9: ML analysis of NMB0992 using a variety of models.p_1, p_2 and p_3 refer to the proportion of sites in categories 1, 2 and 3 respectively. ω refers to the *d*N/*d*S ratios in these categories of sites. p and q are beta estimates. Models M1 and M7 do not allow positively selected sites. Positively selected codons that are bold and underlined represent the codons that are found by both the ML and MP method as being under the influence of positive Darwinian selection.

Model	Estimates of Parameters	ln L	2Δln L	Positively selected codons
M0 (one ratio)	ω1 = 0.3868	-3396.06		None
M1 (neutral)	$p_1 = 0.81$ $\omega_1 = 0.00$ $p_2 = 0.18$ $\omega_2 = 1.00$	-3326.86		Not allowed
M2 (selection)	$p_1 = 0.87$ $\omega_1 = 0.00$ $p_2 = 0.02$ $\omega_2 = 1.00$ $p_3 = 0.09$ $\omega_3 = 3.95$	-3298.77	(M1 *vs* M2) 56 $p < 0.0005$	18, **35**, 44, 50, 62, 64,65 67, 78, 83, 102,149, 176 278 373, 417, **472**
M3 (discrete k=2)	$p_1 = 0.88$ $p_2 = 0.11$ $\omega_1 = 0.00$ $\omega_2 = 3.67$	-3298.84	(M0 *vs* M3) 194 $p < 0.0005$	11, 18, **35**, 44, 50, 51,59, 61, 62, 64,65, 67, 70,74, 78, 83, 94,102, 120,135,137, 139, 143, 149, 176, 200, 202, 240, 254, 261, 278, 338, 351, 372, 373, 374, **382** **383**, 417, 426, 439, 454, **472**, **476**, 491

M7 (beta)	p= 0.00130
	q= 0.00455
		-3327.02		Not allowed

M8 (beta & ω)	p = 0.001 q= 1.81 $p_2 = 0.11$ $\omega_2 = 3.67$	-3298.84	(M7 *vs* M8) 56 $p < 0.0005$	11, 18, **35**, 44, 50, 51,59, 61, 62, 64,65, 67,70,74, 78, 83, 94, 102, 120,135,137, 139, 143, 149, 176, 200, 202, 240, 254, 261, 278, 338, 351, 372, 373, 374, **382** **383**, 417, 426, 439, 454, **472**, **476**, 491

Table 3.10: Results of the ML analysis of NMB2132 using a variety of models.p_1, p_2 and p_3 refer to the proportion of sites in categories 1, 2 and 3 respectively. ω refers to the *d*N/*d*S ratios in these categories of sites. p and q are beta estimates. Models M1 and M7 do not allow positively selected sites. Positively selected codons that are bold and underlined represent the codons that are found by both the ML and MP method as being under the influence of positive Darwinian selection.

Model	Estimate of Parameters	ln L	2Δln L	Positively selected codons
M0 (one ratio)	$\omega = 0.57$	-6827.19		None
M1 (neutral)	$p_1 = 0.58$ $p_2 = 0.42$	-6609.20		Not allowed
	$\omega_1 = 0.00$ $\omega_2 = 1.00$			
M2 (selection)	$p_1 = 0.56703$ $\omega_1 = 0.00$	-6552.92	(M1 *vs* M2) 435	**217**, **222**, **224**, **260**,**264**,**268,276**, 284,
	$p_2 = 0.33670$ $\omega_2 = 1.00$		$p < 0.0005$	**290**, **293**, **305**, **309**, **311**
	$p_3 = 0.09627$ $\omega_3 = 3.97$			
M3 (discrete k=2)	$p_1 = 0.70$ $\omega_1 =0.07$	-6570.05	(M0 *vs* M3) 514	**193**, **217**, **222**, **224**, **260**,**264**,**268,276**,
	$p_2 = 0.30$ $\omega_2 = 2.13$		$p < 0.0005$	284, **289** **290**, **292**, **293**, **305**, **309**,**311**
M7 (beta)	$p = 0.018$ $q = 0.035$	-6612.12		Not allowed
M8 (beta & ω)	$p = 0.020$ $q = 0.03$	-6552.80	(M7 *vs* M8) 118	**193**, **217**, **222**, **224**, **260**,**264**,**268,276**,
	$p_1 = 0.10858$ $\omega_1 = 3.66$			284, **289** **290**, **292**, **293**, **305**, **309**,**311**

models that allow for different dN/dS ratios among evolutionary lineages (Yang 1998) were used (Figures 3.3 and 3.4). In all cases none of the LRTs performed allowed for positive selection on any one lineage exclusively (Table 3.11). This indicates that the results from the first set of LRTs are the result of constant selection on specific sites across all lineages and is not focussed on any particular lineage.

3.3.3.2 Parsimony analysis

A number of sites inferred to be under the influence of positive Darwinian selection using the ML method were not inferred using the parsimony sliding window approach (Tables 3.9 & 3.10). The parsimony method used in this analysis does not consider sites where dS is equal to 0 as having undergone positive selection even if the dN ratio is greater than 0. The ML methods identify situations such as this as positively selected events; therefore, the parsimony method is a more conservative approach. This is the reason why complete agreement between both methods is not observed. In this study, we will treat these sites as being under the influence of positive selection. For NMB0992 the sites inferred to be under positive Darwinian selection have an average ω value of 3.95. An average ω value of 3.12 was found for NMB2132. The ML and MP methods that can detect positive selection are generally in good agreement with one another. Both methods inferred nearly the same set of codons as having undergone positive Darwinian selection for NMB2132 and also for NMB0992 but to a lesser extent (Tables 3.9 & 3.10).

3.3.4 Functionally important residues

A large number of functionally important residues were predicted for both NMB0992 and NMB2132 by ConSeq (Figures 3.5 & 3.6). Obviously the sites that have been shown to be evolving under positive Darwinian selection do not fit into this category, as there must be a degree of variability at such a site before it can be inferred as having undergone positive selection. Interestingly the positively evolving sites inferred by the ML method are nearly all exposed (i.e. are not coiled into the centre of the protein) and therefore probably interact with the host immune system.

Table 3.11: Log likelihood values and parameter estimates using different branch models for NMB0992 and NMB2132. The first column represents the model used and letters in brackets refer to the branch that was labelled see Figure 3.3 and 3.4. Second column lists the number of parameters associated with each model. Log likelihood and transition:transversion ratio are listed in columns 3 and 4 respectively. ω_0 represents the background ω when a particular branch is allowed to vary from the rest of the tree while ω_1 corresponds to that variable ω. The one ratio model assumes that all branches in the tree have a uniform ω value, alternative models that allow for variable ω along predefined branches do not explain the data significantly better in any case and are therefore rejected. This leads to the conclusion that lineage specific positive selection is absent from these datasets. ω values for the freeratio model are not displayed as there are a great many of these.

Model	***p***	**lnL**	***k***	ω_0	ω_1
NMB0992 (n = 20)					
One ratio	39	-3396.06	4.705	0.386	n/a
Two ratio (A)	40	-3396.05	4.705	0.386	0.381
Two ratio (B)	40	-3396.06	4.701	0.386	0.381
Two ratio (C)	40	-3396.06	4.705	0.386	0.411
Two ratio (D)	40	-3396.05	4.699	0.386	0.311
Freeratio	75	-3367.605	4.715	n/a	n/a
NMB2132 (n = 18)					
One ratio	34	-6827.19	2.403	0.57	n/a
Two ratio (A)	35	-6827.17	2.402	0.580	0.523
Two ratio (B)	35	-6827.17	2.401	0.581	0.601
Two ratio (C)	35	-6827.17	2.402	0.580	0.579
Two ratio (D)	35	-6827.17	2.402	0.580	0.502
Freeratio	65	6814.10	2.401	n/a	n/a

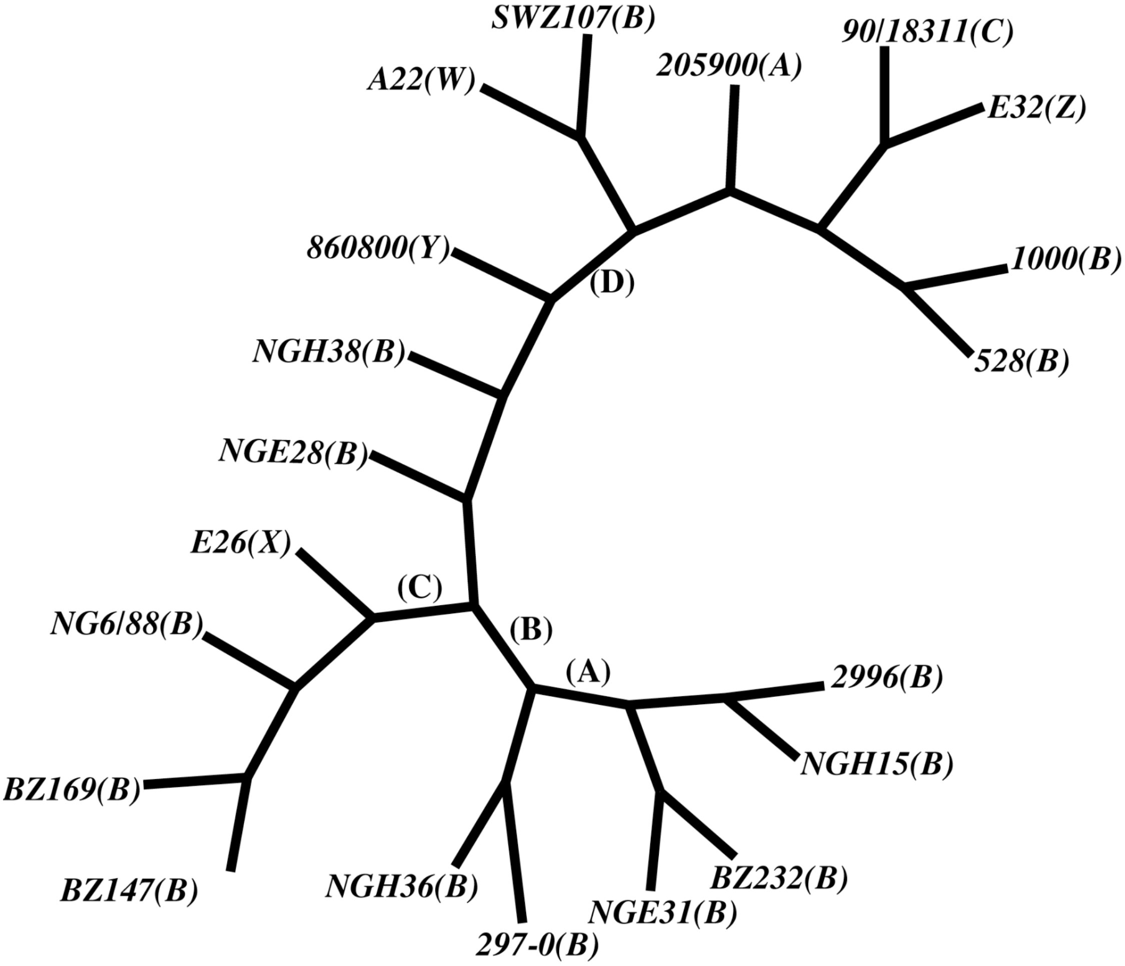


Figure 3.3: Phylogenetic tree for NMB0992 gene inferred by maximum likelihood and Bayesian framework. Branch support values are not displayed as they are relatively low due to lack of phylogenetic signal. This is a result of the sequences used in this analysis sharing a high degree of sequence similarity. To ensure that the tree topology used does not affect the ML analysis, an input tree inferred using the neighbor joining method was used as the input phylogeny. A change in topology did not affect the results reported. Branches labelled A, B, C and D were used in the ML lineage specific analysis. Letters along branches correspond to the lineages tested as having a ω value that differs from the rest of the tree for the branch models of Yang (1998).

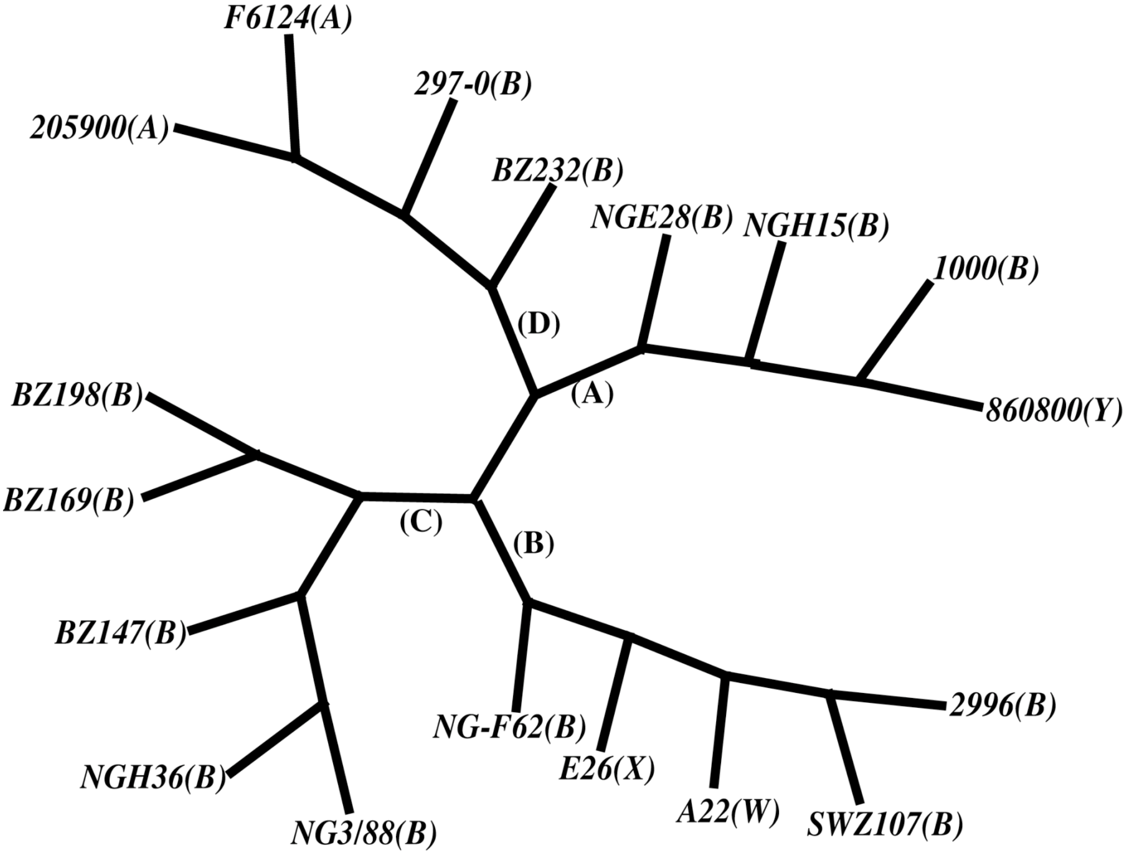


Figure 3.4: Phylogenetic tree for NMB2132 gene inferred by maximum likelihood and Bayesian framework. Branch support values are not displayed as they are relatively low due to lack of phylogenetic signal. This is a result of the sequences used in this analysis sharing a high degree of sequence similarity. To ensure that the tree topology used does not affect the ML analysis, an input tree inferred using the neighbor joining method was used as the input phylogeny. A change in topology did not affect the results reported. Letters along branches correspond to the lineages tested as having a ω value that differs from the rest of the tree for the branch models of Yang (1998).

```
MNKIYRIIWN SALNAWVVVS ELTRNHTKRA SATVETAVLA TLLFATVQAN
eeebbebbbb bbbebbbbbb ebbeebeeeb ebebebbbbb bbbebbbebe
fffssfssss  ssfsss ss fssffsfffs fsfs sssss sss sssfs  50

AFTYSLKKDL TDLTSVGTEE LSFGANGNKV NITSDTKGLN FAKKTAGTNG
bbeeebeeeb eebeeeeeee bebeeeeeeb ebebeeebbe bbeeeeeeee
 sfffsff s   s  f ff  sfs fff fs fs sfffssf ssf ffffff 100

DTTVHLNGIG STLTDRAASI KDVLNAGWNI KGVKTGSTTG QSENVDFVRT
eeebebbbbb bebeeebeeb eebbebbbeb eeeeeeeeee eeeebebbbb
f fsfsssss sfsfffsff  ffssfsssfs ffff f f f ff fsfss s 150

YDTVEFLSAD TKTTTVNVES KDNGKRTEVK IGAKTSVIKE KDGKLVTGKG
bebbebbeee eeebebebeb eeeeeebebe bebeebbbee eeeeebeeee
sfssfssfff fffsfsfsfs fffff sfsf sfsffsssff fffffsfff  200

KGENGSSTDE GEGLVTAKEV IDAVNKAGWR MKTTTANGQT GQADKFETVT
eeeeeeeeee eebebbbeeb bebbbebebe beeeeeeeee eebeebeebe
f ffffffff ffsfsssffs sfsssfsfsf sffffffff  ffsffsffsf 250

SGTNVTFASG KGTTATVSKD DQGNITVKYD VNVGDALNVN QLQNSGWNLD
eeeebebbee eeeebeeeee eeeebbbeee bebbebbebe ebeeeeeebe
fff sfssff  fffsfffff ffffsss ff sfssfssfsf fsffffffsf 300

SKAVAGSSGK VISGNVSPSK GKMDETVNIN AGNNIEITRN GKNIDIATSM
bebbeeeebe bbbbeeeeee eebeeebebe beeebebeee beebebbbbb
sfssffffsf ssssffffff ffsfffsfsf sfffsfs ff sffsfsssss 350

TPQFSSVSLG AGADAPTLSV DDEGALNVGS KDANKPVRIT NVAPGVKEGD
beebbebbbe bbbebeebeb eeeebbebee eeeeeebebb ebbeebeeee
 ffssfsssf sssfsffsfs f   ssfsff f  fffsfss fssffsffff 400

VTNVAQLKGV AQNLNNHIDN VDGNARAGIA QAIATAGLVQ AYLPGKSMMA
beebbebeeb beebeeebee bebebeeebb ebbbebebbb bebeeebbbb
sfssffffsf ssssff fff ffsff sfsf sfffsfsf f sffsfsssss 450

EAGYAIGYSS ISDGGNWIIK GTASGNSRGH FGASASVGYQ W
eebebbbeeb bebbbebebe beeeeeeeee eebeebeebe e
fff sfssss fssfsssfss fssff fffs fsfsssssss
```

Figure 3.5: ConSeq predictions of structure for NMB0992 demonstrated on AF226356, using all homologous genes within this family. The sequence of the query protein is displayed on the first row. The second row lists the predicted burial status of the site (i.e. 'b'-buried versus 'e'-exposed). The third row indicates residues predicted to be structurally and functionally important 's' and 'f' respectively. Sites inferred as evolving under the influence of positive selection are in bold type and underlined, the majority of such sites are in exposed regions of the protein.

```
MFKRSVIAMA CIVALSACGG GGGGGSPDVK SADTLSKPAA PVVTEDVGEE
eeeebbbbbb bbbbbbbbbe eeeeeeebee beebeeebbe bbbeeebeeb
ffffssssss ss ssssssf fffffffsff sffsfffssf ss fffsffs 50

VLPKEKKDEE AVSGAPQADT TQQDATAGKG GQDMAAVSAE NTGNGGAATT
beeeeeeeee eeebeeeeee ebeeeeebee bbbbbeeeee eeeeeeeeee
sfffff  ff    sfff  f  s    f f  sssss ffff ffff f fff 100

DNPENKDEGP QNDMPQNAAD TDSSTPNHTP APNMPTRDMG NQAPDAGESA
eeeeeeeeee eeeebeeeee eeeeeeeeee bbbeeeeeeb eeeeeeeeee
 f ff  f f fff sf fff  ffffff ff s    f ffs  f fff fff 150

QPANQPDMAN AADGMQGDDP SAGEENAGNT ADQAANQAEN NQVGGSQNPA
ebeebbebbe ebeeeeeeee eebeeebeeb eeeeeeeeee eeeeeeeeee
fsffssf sf f ffffff f  fsfffs f   fff fff   ff       f 200

SSTNPNATNG GSDFGRINVA NGIKLDSGSE NVTLTHCKDK VCDRDDFLDE
eeeeeeeeeb bebebbebbe bebeeeebbb eebeeeeeee ebbeeeeeee
    f f  s sf             f f ss ffsf   f     sfff   f 250

EAPPKSEFEK LSDEEKINKY KKDEQRELEN NNFVGLVADR VEKNGTNKYV
eebeebeeee ebeebeeeee eeebbebbbe ebeeeeeeeb beebeeeeee
ffs  sf        sf        s   ss       f f  s    fff    300

IIYKDKSASS SSARFRRSAR SRRSLPAEMP LIPVNQADTL IVDGEAVSLT
eebbebeebb ebeeebeeeb ebbebeebeb bbbebebbbb bbeebebbbe
   s sff s   fffs  fs fssfsffsfs sssfsfssss ssffsfsss  350

GHSGNIFAPE GNYRYLTYGA EKLSGGSYAL SVQGEPAKGE MLAGTAVYNG
eeebebbbeb beebeebbbb bebeeeebee ebbbbbbbbe bebbbbebee
fffsfsssfs sffs fssss s sffffsff fssss sssf sfssssf ff 400

EVLHFHMENG RPSPSGGRFA AKVDFGSKSV DGIIDSGDDL HMGTQKFKAV
eeeeeeebeb bbebeeebee bebbbbbbee beebeeebeb bbeeeebebe
ff     sfs ssfsfffsff sfssssssff sffs ffsfs  sffffsfsf 450

IDGNGFKGTW TENGGGDVSG RFYGPAGEEV AGKYSYRPTD AEKGGFG
eeeeeeeeee bebebeeeee bbeeeeeeee eeeeeebbbb beeeeee
fffff ffff s sfsfffff ssffffffff ffffffssss sffffff
```

Figure 3.6: ConSeq predictions of structure for NMB2132 demonstrated on AF226419, using all homologous genes within this family. The sequence of the query protein is displayed on the first row. The second row lists the predicted burial status of the site (i.e. 'b'-buried versus 'e'-exposed). The third row indicates residues predicted to be structurally and functionally important 's' and 'f' respectively. Sites inferred as evolving under the influence of positive selection are in bold type and underlined, the majority of such sites are in exposed regions of the protein.

3.4 Discussion

3.4.1 OMP85 homologues

The methods used to detect positive selection in this analysis included ML and MP. Each method has it limitations but in general the ML method has been considered the most robust criterion for detecting adaptive evolution in protein coding genes (Yang 2002). Recent criticisms (Suzuki and Gojobori 1999; Suzuki and Nei 2001; Suzuki and Nei 2004) suggest that the ML method produces many false positives. To account for this problem I also utilised a MP approach. High levels of recombination can also affect ML analysis (Anisimova *et al.* 2003) but no evidence of recombination was found in the data set used in this study. This is in agreement with the initial characterisation of this gene family which found no hypervariable or recombining regions (Voulhoux *et al.* 2003).

In this study, positive selection has been shown to have acted on a number of amino acid sites the majority of which are located in surface exposed loops, thus revealing the evolutionary processes acting on OMP85. ML and MP analysis of selective pressures found that approximately 3% of all codon sites have been under the influence of positive Darwinian selection. Mapping these sites onto a two dimensional model confirmed that the majority of sites were found in surface exposed loops. Surface exposed loops are predicted to be recognised by host responses (Urwin *et al.* 2002) and are therefore under the greatest pressure for change in an attempt to stay one step ahead of their host/environment in what can be described as an 'arms race' (Jiggins *et al.* 2002). Transmembrane regions of proteins are likely to be highly conserved (Jiggins *et al.* 2002). This analysis failed to locate any positively evolving sites in transmembrane regions this is probably due to important structural constraints in anchoring OMP85 in the OM. Until a detailed OMP85 tertiary structure is described it is not possible to ascertain the significance of the 14 positively evolving sites. It is reasonable to assume they may be involved in foreign antigen recognition by host responses therefore it may be advantageous for pathogenic bacteria to be able to alter the confirmation of these loops to increase antigenic diversity. One positively evolving site was found in the periplasmic NH_2-terminal domain, again it is impossible to draw any definitive conclusions from this result in the absence of a tertiary structure. However possible advantages may include an increased ability to fold membrane proteins.

3.4.2 Vaccine candidate genes

The development of a vaccine against heterologous *N. meningitidis* B strains is an ongoing process. Despite more than 20 years of research (Frasch 1989; Bjune *et al.* 1991; Sierra *et al.* 1991) vaccines to protect against heterologous strains have yet to be developed. Candidate vaccine designs including the purification of the polysaccharide capsule is not an option as it may lead to autoimmunity. Researchers are instead trying to locate highly conserved membrane proteins that elicit a host immune response. This analysis has found that two such vaccine candidates display evidence of positive selection. This finding may have implications for vaccine design as a protein that has exhibited an ability to undergo selection in the past may do so again. While we can only speculate, it is plausible that further change at important antigen binding sites may make any vaccines developed from these proteins obsolete.

In this analysis, LRTs were used to model variable selective pressures across sites and lineages for a variety of vaccine candidates that are currently being tested. Recently concerns have been raised regarding the frequent false positive results inferred using this ML approach (Suzuki and Nei 2001; Suzuki and Nei 2004). In an effort to add further validity to the ML results, a parsimony analysis was also performed. According to the ML analysis for selection within NMB2132, 16 sites are under the influence of adaptive evolution, the parsimony method infers 15 of these sites also. There were some discrepancies between the two methods when NMB0992 was analysed as the ML method inferred 45 sites compared to the parsimony method which, only inferred 5 of these sites. Therefore we conclude that there is excellent evidence that 15 sites within NMB2132 have undergone an adaptive event while there is corroborated evidence to suggest that at least 5 sites within NMB0992 have also undergone an adaptive event. Furthermore the majority of these sites are within exposed regions therefore are likely to be in direct contact with host immune responses.

High levels of recombination can affect ML analysis (Anisimova *et al.* 2003) as can saturation of synonymous sites. These concerns were carefully considered yet in all seven gene families examined no signatures of major recombination events or saturation of synonymous sites were observed. Previously these seven proteins have been shown to have extremely low levels of recombination (Pizza *et al.* 2000).

Previous vaccine studies have concluded that vaccines targeted against epitopes which consist of negatively selected sites protect more efficiently than those directed against epitopes which contain positively selected sites. Using this logic we suggest that of the seven *Neisseria* vaccine candidates analysed here careful examination needs to be given to those that exhibit evidence of positive selection. Among a suite of seven proteins that are known to be highly expressed, known to be surface exposed and known to be capable of eliciting a strong bactericidal immune response, two showed evidence of having undergone positive selection. The other five proteins do not display any historical evidence of positive selection for change. The fact that evidence of positive selection was found in some of these proteins is not particularly surprising given that other membrane proteins have been shown to evolve positively (Smith *et al.* 1995; Jiggins *et al.* 2002; Andrews and Gojobori 2004). Intuitively we would expect these proteins to be under the greatest pressure for change in an effort to remain ahead of host immune responses in the host-pathogen “arms race”. What are the implications of this kind of analysis? At this current moment in time, all these proteins still elicit a bactericidal immune response. However, if vaccines target epitopes that contain positively selected sites such as those found in NMB0992 and NMB2132 it is feasible that future amino acid altering substitutions at these sites may render such vaccines obsolete. From this analysis, we would suggest that this kind of event is much more likely in those proteins that have historically shown an ability to evolve under positive selection than in those proteins where positive selection has not been a feature. We therefore propose that an analysis of historical adaptive evolution seems to be a sensible precautionary measure prior to the expensive process of developing a vaccine. Following this line of thought the remaining five vaccine targets are ideal vaccine candidates due to the absence of positive Darwinian selection and any vaccines developed against NMB0992 and NMB2132 should avoid targeting the regions we have inferred to be under the influence of positive selection.

Chapter 4 Is there a phylogeny among four completely sequenced *Neisseria* isolates?

4.1 Introduction

The principle objective of a phylogenetic study is to contribute to the discovery of the true species phylogeny of life underlying biological diversity (Penny and Hendy 1986). When different data sets from the same organisms are analysed using robust phylogenetic reconstruction methods, they are expected to converge on the true phylogeny for their species (Miyamoto and Fitch 1995). Therefore, when there is a high degree of congruence between independent data sets we should have a good estimate of the true phylogeny for a group (Lanyon 1993). There are many problems associated with the analysis of multiple independent data sets. These problems have led to debate regarding the optimal strategy of analysis of such datasets (Bull *et al.* 1993; De Queiroz 1993). The last fifteen years have witnessed a large body of published work pertaining to a number of distinct solutions including real and simulated data in an effort to demonstrate the virtues of one solution over another. Today we are no nearer a universal view on how best to proceed with such data. Some argue that all data should be combined before a phylogenetic analysis (character congruence (Kluge 1989; Eernisse and Kluge 1993)) while others argue that the independent trees obtained for each dataset should be combined using consensus techniques (taxonomic congruence (Kluge 1989)) (Figure 4.1). The debate regarding the advantages and pitfalls of both taxonomic and character congruence is a contentious but important controversy in modern systematics (Miyamoto and Fitch 1995). Different data may evolve under different evolutionary processes, resulting in heterogeneity of the individual data partitions that make up the overall dataset. How best to deal with this heterogeneity is the question that still needs to be answered, but there are three broad proposals. The first is to always combine data (Barrett *et al.* 1991), the second is to never combine data (Miyamoto and Fitch 1995) and a third intermediate approach which is to only combine data after careful examination (Bull *et al.* 1993).

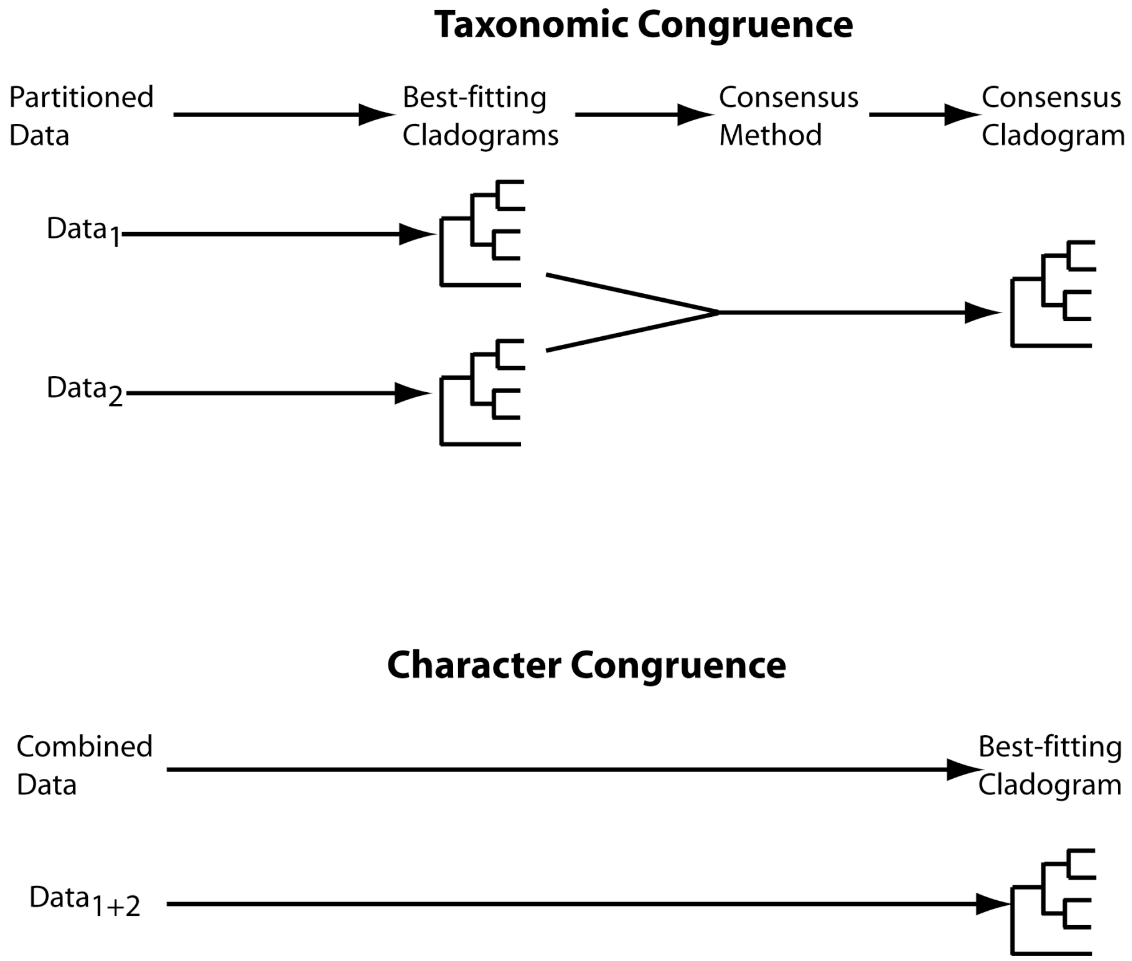


Figure 4.1: Character congruence and total congruence, compared in terms of the steps involved in analysing partitioned or combined data. The former approach employs a consensus method. Redrawn from Eernisse and Kluge (1993).

Advocates of character congruence (CC) argue that combining data is justified as it increases the informativeness of the character data used in the analysis. As an example the addition of two characters, the first of which resolves nodes close to the tips while the second is useful for resolving the basal branching, should substantially improve the resolution of the entire end tree. Also if weak but true signal is present in a number of data sets it is hoped that combining the data will have an additive affect and should express itself in the recovered phylogeny (Barrett *et al.* 1991; Flook *et al.* 1999). Justification for CC has been sought from first principles, in other words phylogenetic hypotheses must be selected based on all available information (Kluge 1989; Barrett *et al.* 1991). Alternatively, opponents of CC point to the fact that support for true phylogenetic groupings may be diluted by systematic or random errors from unreliable characters. In a worst-case scenario a combined analysis will give an incorrect estimate of phylogeny with increasing certainty as the size of the data increases (Bull *et al.* 1993).

As more sequence data has become available, the issue of horizontal gene transfer (HGT) and the problems it creates when trying to elucidate taxonomic relationships has been highlighted (Daubin *et al.* 2003). HGT will result in incongruence between the gene tree and the species tree (Wiens 1998). The phenomenon of HGT and others such as hidden paralogy has provided a very strong argument against combining such data sets (De Queiroz 1993) and has led to the view that genes with different histories should never be combined to produce a single representative tree (Bull *et al.* 1993). Instead, a measure of the degree of incongruence should be taken and only after careful examination should data be combined (Bull *et al.* 1993; De Queiroz 1993).

Using both TC and CC methodology this analysis wishes to investigate if there is a phylogeny among *Neisseria* isolates for whom complete genomes are available. The *Neisseria* dataset presented here is interesting as HGT occurs continually in *Neisseria* populations (Maiden and Feavers 1995). High rates of HGT in *Neisseria* has led to the blurring of gene-pool boundaries and has subsequently led to the concept of a global gene pool for the *Neisseria* (Maiden *et al.* 1996). This global gene pool makes inferring true phylogenetic relationships difficult. If the majority of genes in the *Neisseria* dataset support a single topology then one would expect that any method used to analyses this dataset will return this topology with high support values for internal branches. Alternatively, if no single topology is supported by a large majority of the individual genes than we would expect that any method used to summarise this dataset should return a poorly supported phylogenetic tree. Two *Neisseria* datasets are used in this analysis; the first dataset contains genes that have definite phylogenetic signal while the second consists of genes that have no definite signal according to the appropriate statistical tests. Genes may lack phylogenetic signal if they are very closely related or if the degree of substitution has been so great that the variable positions in an alignment are effectively randomised and this is true for approximately 70% of the genes used in this study. It has been suggested that concatenating such sequences will reveal the true phylogeny as the resultant concatenated alignment will contain more information than any single gene (Flook *et al.* 1999). Similarly concatenating genes that have definite phylogenetic signal should also have an additive effect assuming that the individual genes are congruent with one another the majority of the time (Bull *et al.* 1993). Simulation studies that investigate the frequency of false positive results with regard to three different phylogenetic reconstruction methods are also examined.

4.2 Materials and Methods

4.2.1 Sequence data

The 1,190 single gene *Neisseria* families from Chapter 2 were used in this analysis. See section 2.2.2 for a complete description of the methods used to locate and align these gene families.

4.2.2 Tests for phylogenetic signal and bootstrap analysis

From the 1,190 single gene families an analysis was carried out to determine which genes had phylogenetic signal according to the appropriate statistical tests. These genes will act as the first dataset in this study ("signal" dataset). To perform this step of the analysis a permutation tail probability test (PTP test) (Archie 1989; Faith and Cranston 1991) was used. The PTP test is a test for congruence across characters; this congruence may be due to phylogeny or even stochastic effects such as similar base composition. The PTP test proceeds by selecting all columns of a given datamatrix (alignment), independently shuffling each column and reassigning it to a different species at random. Manipulating the data in this way breaks all compatibility between characters within the matrix but maintains the dataset size and also the nucleotide composition at each site. A distribution of goodness-of-fit measures such as parsimony scores among these permuted data sets are compared to the value from the original data. If the score from the real data lies far enough into the tail of the distribution scores then there is structure in the data (Felsenstein 2004). In total, 390 families were shown to have character congruence at a level that is significantly better than random chance according to the PTP test.

There have been criticisms that the PTP test too readily detects significant phylogenetic structure (Slowinski and Crother 1998; Wilkinson *et al.* 2002), furthermore the PTP test only looks for deviation from random levels of congruence which may be due to stochastic effects and not necessarily phylogenetic signal. Therefore for completeness, a second test of phylogenetic structure was performed. For the 390 genes, phylogenetic relationships using the maximum likelihood (ML) framework were inferred with the best fitting model of sequence evolution for that gene as determined by MODELTEST 3.06 (Posada and Crandall 1998). The ML derived topology was then compared to the two alternative topologies (there are 3 possible topologies for 4 taxa) using the Shimodaira Hasegawa test (SH test) (Shimodaira and Hasegawa 1999). The SH test is a resampling method that approximately corrects for testing multiple trees. It proceeds by making R bootstrap samples of the N sites present in the given alignment. The total log likelihood is computed for each sample. The mean score of each tree across all R bootstrap samples is subtracted from the sum of the resampled log likelihoods. This “normalising” has the effect of adjusting all trees so that their log likelihoods have the same expectation. Therefore, if the total log likelihood of the *i*th and *j*th bootstrap sample is ℓ_{ij} the normalised value R_{ij} is computed as

$$R_{ij} = \ell_{ij} \frac{1}{R} \sum_{k=1}^{R} \ell_{ik}$$

Taking the *j*th bootstrap replicate and computing by how much the normalised value of the *i*th tree is below the maximum across all trees for that replicate is given by

$$S_{ij} = (\max R_{kj}) - R_{ij}$$

Now for each tree *i*, the tail probability is given by the fraction of the bootstrap replicates in which S_{ij} is less than the actual difference between the maximum likelihood and the log likelihood L_i of that tree. Therefore, trees with tail probabilities above a specific target (normally 0.05) cannot be rejected. If more than one of the trees meets this target then we cannot say one topology fits the data significantly better than a competing phylogenetic hypothesis. In total, 344 datasets were found to pass the PTP test and subsequently found to support a single topology significantly better than any of the alternative two competing topologies. These 344 genes make up the "signal" dataset. The PTP and SH test were both performed using PAUP* 4.0b10 (Swofford 1998). The remaining 846 genes were deemed to lack robust phylogenetic signal and these make up the "no signal" dataset.

The "signal" and "no signal" datasets were concatenated separately to yield two alignments of 402,813 and 789,375 nucleotide positions in length respectively. The best fitting models of sequence evolution for these combined datasets were found using MODELTEST 3.06 (Posada and Crandall 1998). In order to assess support levels for internal branches, the dataset was resampled using the bootstrap technique (Felsenstein 1985). ML, neighbor joining (NJ) and parsimony trees were constructed from distance matrices based on these resampled datasets and summarised using a majority rule consensus tree.

4.2.3 Simulations

4.2.3.1 Overcredibility of branch supports obtained by bootstrap resampling

The nonparametric bootstrap was first applied to phylogenetics by Felsenstein (1985). It uses a data resampling technique to assess support for particular phylogenetic trees. The bootstrap is the most widely used statistical assessment of inferred phylogenetic relationships (Cummings 2003) although its use is somewhat controversial (Sanderson 1995). An investigation into possible false positive branch supports from bootstrapping as inferred by the ML framework has been performed in the past (Suzuki *et al.* 2002; Cummings *et al.* 2003; Erixon *et al.* 2003). To investigate these potential problems with respect to the *Neisseria* data, I wished to examine the correlation between sequence length and the number of false positive results reported for ML, NJ and parsimony bootstrap analyses. Four datasets each containing 100 replicates were generated for this purpose. Each replicate contained 4 taxa and consisted of 3 equal length partitions. The three partitions were generated along the following topologies ((a^1,b^1),(c^1,d^1)) (Figure 4.2A), ((a^2,c^2),(b^2,d^2)) (Figure 4.2B) and ((a^3,d^3),(b^3,c^3)) (Figure 4.2C) using a Monte Carlo technique as implemented in Seq-Gen 1.2.5 (Rambaut and Grassly 1997). Interior branch lengths (b_I) were set to 0.02 substitutions per site, exterior branch lengths (b_E) were set to 0.035 substitutions per site for *N. meningitidis* serogroups A, B, C and 0.06 substitutions per site for *N. gonorrohoeae*. Branch lengths were selected based loosely on branch lengths observed from the concatenated alignment of the 344 genes from the “signal” dataset. All four generated datasets were similar except that the size of the individual partitions within them varied (i.e. 500, 5,000, 10,000 and 100,000 nucleotide sites). Sequences a^1, a^2 and a^3 were concatenated to give a single sequence A, this was also done for the other generated sequences to yield sequence B, C and D, this procedure was followed for the four data sets of varying size. In total four concatenated alignments were created that were 1,500, 15,000, 30,000 and 300,000 nucleotides in length. The topology of each concatenated alignment should resemble a bush (Fig 4.2D). However one topology is generally chosen due to stochastic error, despite this, it should not receive high bootstrap support (Suzuki *et al.* 2002). Cases when bootstrap scores are high can be considered false positives (scores above 80% were considered to be false positives). Suzuki *et al* (2002) only considered scores above 95% as evidence of a false positive result. This score was considered to be too strict as previous studies that combine data have used lower bootstrap support values as a basis for justifying certain relationships (Baldauf *et al.* 2000; Brown *et al.* 2001; Brochier *et al.* 2002; Brochier *et al.* 2004).

4.2.3.2 Effect of additional sequence data on bootstrap supports

The goal of this second simulation was to determine which phylogenetic reconstruction method will converge on a high bootstrap score the quickest. This was achieved through the

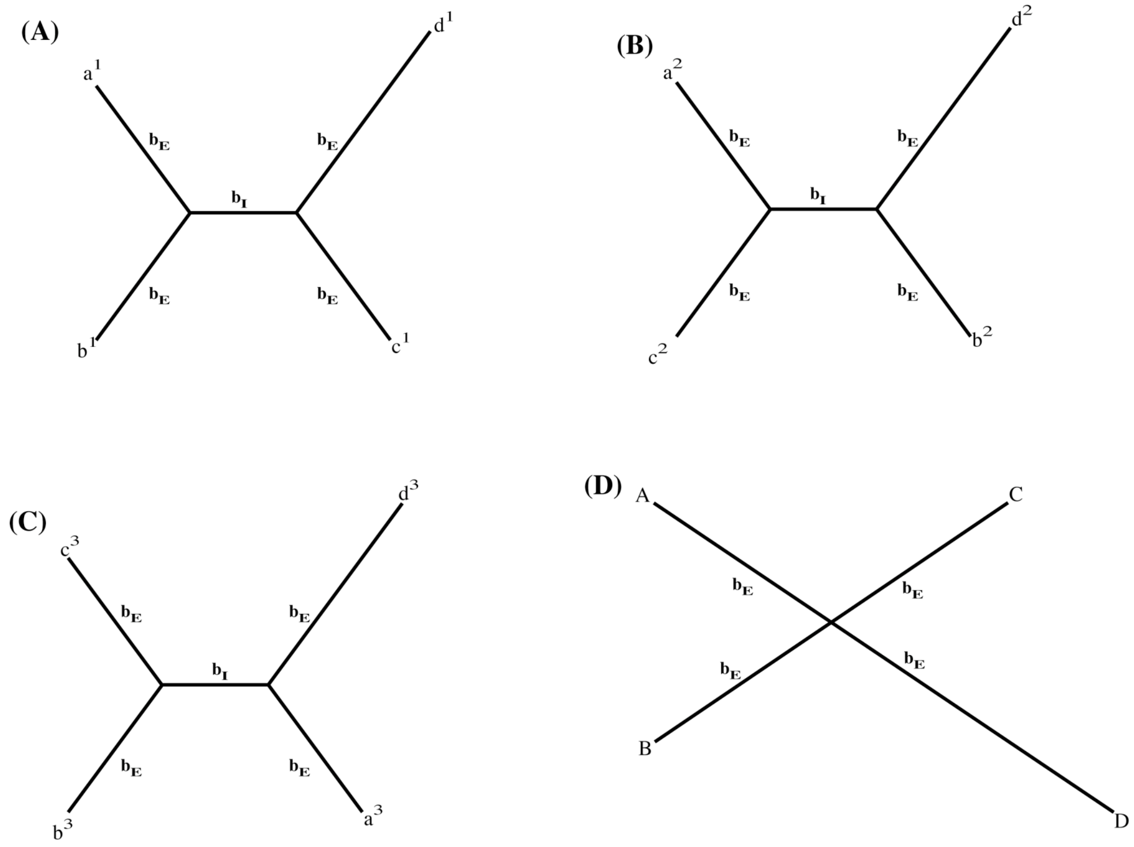


Figure 4.2: Model topologies utilised in generating concatenated sequences. b_E and b_I represent exterior and interior branch lengths respectively. Exterior branch lengths of a, b and c are all equal, the branch length for taxon d is slightly longer than others (as is the case for *N. gonorrohoeae* in the real dataset). No interior branch exists in the star phylogeny depicted in D.

successive addition of sequence data to an alignment. Bootstrap replicates were created and phylogenies based on ML, NJ and parsimony reconstruction methods were used to infer phylogenetic relationships. The rationale behind this simulation is that as more data is added to the alignment stochastic errors should dissipate and ideally unrepresentative bootstrap supports should not be observed. Initially an alignment of 10 nucleotides positions was created, 4 of the positions support topology I, 3 positions support topology II and 3 positions support topology III (Figure 4.3). The length of the simulated nucleotide alignment is incrementally increased with the frequency of sites supporting the three possible topologies remaining constant. ML, NJ and parsimony bootstrap analyses were performed on each of the alignments. Sequence length and bootstrap support values were recorded.

4.2.3.3 Spectral analysis

A spectral analysis of the concatenated data sets was performed using the Spectrum software (Charleston 1998). The Sequence alignment is analysed site by site categorising observed nucleotide distribution patterns at each position and scoring the occurrence of splits found within the alignment. At each position, data is split into two subsets, those with the same nucleotide at that site and those with a different nucleotide at that site. The frequency of a split is the sum of occurrences of that split in the data divided by the total number of columns in the data set. Frequencies are specifically called support for a particular split (Lento *et al.* 1995). If the data were extremely homogenous then there would only be evidence for one set of mutually compatible phylogenetic relationships in any given data set. This scenario is very unlikely in the real world and it is usual to observe contradictory information at a single site. Therefore, as well as determining support values for specific splits it is also possible to quantify the amount of conflict against each split. The conflict value for a split is the sum of all other splits that contradict the portioning of taxa in the first split. Because a split may be incompatible with many other splits, it is feasible that a split may have a degree of conflict that is greater than the support it receives. Using this strategy, it is possible to examine the data and determine a measure of support independent of a phylogenetic tree.

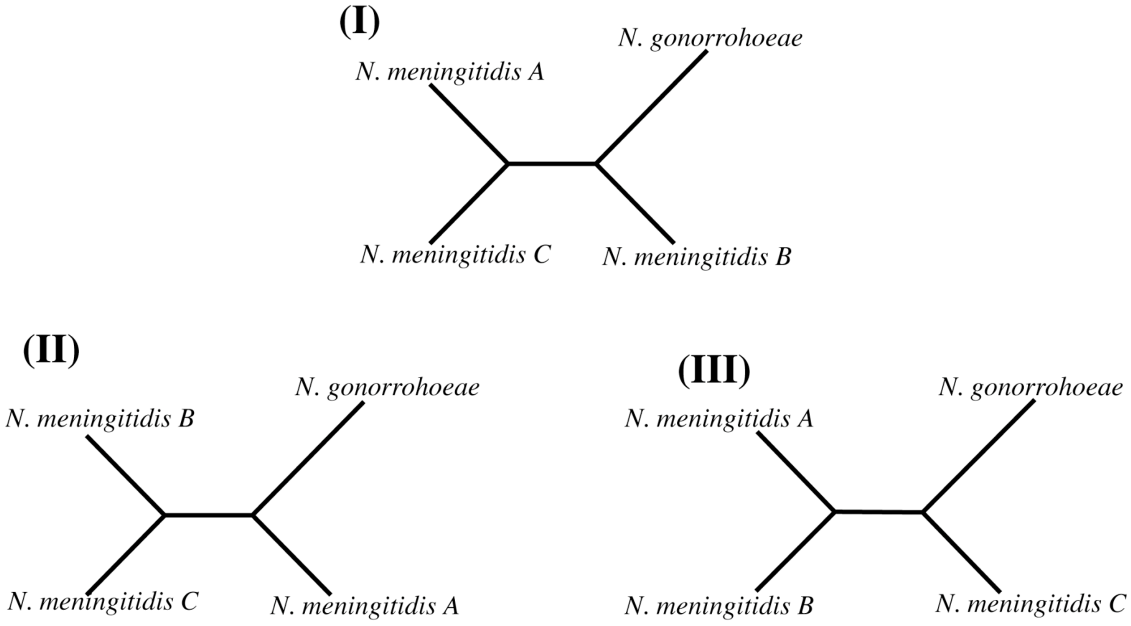


Figure 4.3: The three possible tree topologies for the *Neisseria* four taxon dataset.

4.3 Results

4.3.1 Genes with phylogenetic signal

From the 344 genes deemed to have phylogenetic signal according to the PTP and SH test, it was found that 128 (37%) trees support the sister group relationship of serogroups A and C (topology I), 113 (33%) trees support the sister relationship of serogroups B and C (topology II) and 103 (30%) trees support the sister relationship of serogroups A and B (topology III). From these results, it is obvious that there is a high degree of incongruence within this dataset as each of the three possible topologies are supported approximately equally. Concatenating the "signal" dataset into a single alignment yields a 402,813-nucleotide alignment. There are three possible unrooted topologies for a four-taxon dataset (Figure 4.3). A parsimony bootstrap analysis of the parsimony informative sites from the concatenated alignment gives 100% bootstrap support (BS) for topology I. A distance bootstrap analysis (distances based on general time reversible (GTR) model (Lanave *et al.* 1984) and summarised using NJ trees) supports topology I with 95% BS. MODELTEST 3.06 tested 56 possible models of sequence evolution and found that the GTR model of substitution with sites distributed along a gamma distribution and a proportion of invariable sites (GTR+I+G) best describes the concatenated alignment. This model of evolution as well as the estimated parameters were used as the default parameters for an analysis using ML as the optimality criterion with the robustness of phylogenetic relationships being evaluated using the bootstrap resampling procedure. From this analysis the grouping of serogroups A and C (topology I) is supported with 80% BS.

Closer examination of the concatenated alignment revealed that of the 402,813-nucleotide positions only 7,918 (~2%) are informative. This is not surprising given the low level of sequence divergence between the sequences. Using the retention index (Clyde and Fisher 1997) implemented in PAUP* 4.0b10 it was discovered that of the informative sites, 2,851 (36%) support topology I, 2,612 (33%) support topology II and 2,455 (31%) support topology III. As with the individual underlying 344 trees the distribution of informative characters supporting the three alternative topologies is approximately equal. Informative characters that support the sister relationship of serogroups A and C (topology I) are represented marginally more than the alternative two topologies. Yet, when the entire concatenated dataset is analysed this grouping always receives a high bootstrap support value regardless of which phylogenetic reconstruction method is used (i.e. ML, NJ or parsimony).

4.3.2 Genes with no phylogenetic signal

The concatenated alignment of the “no signal” dataset, which contained 846 *Neisseria* single gene families was 789,375 nucleotide positions in length. An analysis of parsimony informative sites exclusively, using parsimony as the optimality criterion with the robustness of phylogenetic relationships being evaluated using the bootstrap resampling procedure supports the sister relationship of serogroups A and B (topology III) with 90% BS. A similar analysis of pairwise distances of bootstrap replicates (distances based on the GTR model) and summarised using NJ trees again supports topology III this time with 100% BS. MODELTEST 3.04 finds that the GTR+I+G model best describes the “no signal” concatenated alignment. This model of evolution as well as the estimated parameters were used as the default parameters for a ML bootstrap analysis. From the ML analysis the grouping of serogroups A and B (topology III) was supported with 100% BS. Examination of the “no signal” concatenated alignment showed that of the 789,375-nucleotide positions, only 11,354 (~1.4%) are informative, this percentage is slightly less than what is observed in the “signal” dataset. Of the parsimony informative sites 4,087 (~36%) support topology III, 3,974 (~35%) support topology I and 3,293 (~29%) support topology II. As with the “signal” concatenated alignment, the grouping that has a slight majority of supporting characters (A and B in this case) is always chosen with extremely high BS.

4.3.3 Simulated results

4.3.3.1 Overcredibility of branch supports obtained by bootstrap resampling

In an effort to measure the rate of possible false positive results that may be obtained using a bootstrap analysis I performed a trivial simulation by constructing alignments of varying lengths 100 times. Four sequence lengths were chosen: 1,500, 15,000, 30,000, and 300,000 nucleotide positions. Each alignment consists of three equal length partitions. Each partition supports one of three possible topologies (Figure 4.2). The manner in which these sequences were evolved is over simplistic and in reality probably does not mirror the real evolutionary pressures acting on the *Neisseria* dataset. The advantage of this approach is that although we do not know what the final tree will look like we expect that the bootstrap supports for this tree should be low. Summarising the results we find that approximately 20-25% of trees are actually false positives as they receive a bootstrap score greater than 80% (Table 4.1). A noteworthy finding from the observed false positives is that in all cases they are the result of a small number of informative sites supporting one topology over the others, this observation is similar to results observed from the real data above. For example the frequency of informative sites supporting topologies I, II and III should be approximately equal (i.e. 33.33% for each), however when the false positive results are carefully examined it was found that in all cases the favoured topology was the result of a slight majority of supporting informative characters (Table 4.2). As an example, Table 4.2 illustrates the 18 false positive results from the ML bootstrap simulation performed on a sequence containing 300,000 nucleotides in length. If the percentages of characters supporting one of the three possible topologies are examined it is obvious that the trees that receive inflated BS are the result of a slight majority of characters supporting that topology. Some extreme cases include false positive number 15 and 18 (Table 4.2), in both cases the difference in proportion of characters supporting tree 3 and 2 are 0.004 and 0.002 respectively, yet tree 3 receives a BS value greater than 80% in both cases.

4.3.3.2 Effect of sequence length on bootstrap support values

Sequences alignments of varying lengths were created. All positions were parsimony informative and in all cases 40% of the characters support topology I, the remaining characters support topology II and III equally. In this scenario, topology I is supported marginally better than II and III. When the alignment length is gradually increased (relative supporting character frequencies are kept constant i.e. 40%, 30%, 30%) we find the support

Table 4.1: Results of NJ and ML bootstrap analysis of simulated alignments. All alignments contain three equal length partitions each supporting an alternative topology. One hundred replications were performed for each alignment in some cases topologies were unresolved, as the consensus of bootstrap supports did not find one tree with a score greater than 50%. Numbers in columns correspond to the number of times a particular tree was favoured, numbers in brackets represent false positives. False positives are defined as topologies that receive a bootstrap support greater than 80% for its internal branch. Columns with FP heading summarise the number of false positives found for each simulation.

NJ Bootstrap					**ML Bootstrap**				**Parsimony Bootstrap**			
Length	I	II	III	**FP**	I	II	III	**FP**	I	II	III	**FP**
1500	27 (7)	21 (8)	37 (10)	**25**	27 (6)	18 (6)	37 (9)	**21**	32 (6)	29 (4)	20 (6)	**16**
15000	34 (8)	22 (6)	31 (8)	**22**	37 (8)	21 (7)	27 (7)	**22**	33 (9)	35 (7)	7 (7)	**23**
30000	29 (6)	28 (6)	30 (7)	**19**	25 (9)	25 (5)	31 (7)	**21**	32 (4)	29 (6)	26 (8)	**18**
300000	26 (5)	26 (8)	27 (10)	**23**	32 (5)	22 (5)	28 (8)	**18**	26 (4)	27 (6)	30 (8)	**18**

Table 4.2: False positive results from the ML bootstrap simulation for sequences 300,000 nucleotides in length. Trees with a bootstrap support greater than 80% are considered false positives; in all, there were 18 for this particular simulation (Table 4.1). The frequency columns equate to the proportion of nucleotides that support that particular tree. Bold underlined-numbers detail which tree is reported by the majority consensus, these trees always have a slight majority of nucleotides supporting them, note how in some case this majority is almost negligible

False positive	Tree 1	Tree 2	Tree 3
	Frequency	Frequency	Frequency
1	**<u>0.342</u>**	0.330	0.328
2	**<u>0.343</u>**	0.326	0.331
3	0.333	0.319	**<u>0.349</u>**
4	0.332	**<u>0.341</u>**	0.327
5	**<u>0.343</u>**	0.326	0.331
6	0.321	**<u>0.348</u>**	0.331
7	0.330	**<u>0.341</u>**	0.328
8	**<u>0.345</u>**	0.328	0.328
9	0.330	0.328	**<u>0.343</u>**
10	0.330	**<u>0.344</u>**	0.326
11	**<u>0.345</u>**	0.332	0.323
12	0.332	0.328	**<u>0.340</u>**
13	0.328	0.331	**<u>0.340</u>**
14	0.329	0.323	**<u>0.348</u>**
15	0.328	0.334	**<u>0.338</u>**
16	0.332	0.326	**<u>0.342</u>**
17	0.330	0.326	**<u>0.343</u>**
18	0.320	0.339	**<u>0.341</u>**

from parsimony and NJ methodologies is approximately 50%, but this exceed 90% when the sequence is lengthened to 100 nucleotide positions and reaches 100% at approximately 250 nucleotide positions. An increase in sequence length does not effect the bootstrap supports to the same degree when the ML criterion is used but does so eventually when the sequence length is increased above 3,000 nucleotide positions (Figure 4.4). These results indicate that the parsimony and NJ methodologies may lend high support values to particular topologies even though the underlying data would not appear to be as clear-cut; furthermore, an increase in sequence length (i.e. more data) seems to exacerbate the problem. Based on this simulation the ML methodology does not appear to suffer from the convergence problem to the same degree as the other two reconstruction methods tested.

4.3.3.2 Spectral analysis

With 4 taxa, the number of splits is 8 (2^{4-1}). Examining the "signal" concatenated data there are 3 splits that are supported by some evidence, however the degree of conflict is greater than the degree of support (Table 4.3). Similarly, when the "nosignal" concatenated alignment was examined, 3 splits are supported but the degree of conflict outweighs the measure of support (Table 4.3). These results are not surprising, as previous analysis investigation of the data has shown it to be extremely heterogeneous. (i.e. the 3 possible topologies are supported by a near equal proportion of characters for both alignments. Also, the individual trees for the "signal" datas*et al*so show a near equal distribution amongst the three possible topologies). These results are in stark contrast to the high support values inferred from the bootstrap resampling technique. Lento *et al* (1995) have stated that the spectral signals (this is determined by subtracting the degree of conflict from support for a split) can be reasonable predictors of BS values. This may be true when the data is homogenous or "clean". The *Neisseria* data is not clean and the spectral signals do not correlate with BS values. Instead, they accurately present the underlying signal, which is extremely heterogeneous, and therefore provide a better understanding of the data than the bootstrap analysis alone.

Table 4.3: Results of the spectral analysis. Three splits are supported for both the "signal" and "nosignal" concatenated alignments. Note in all cases how the conflict for these splits exceeds the support. These results are indicative of heterogeneous data. Taxon A, B and C represent *N. meningitidis* serogroups A, B and C respectively.

Split	"signal" data		"nosignal" data	
	Support	Conflict	Support	Conflict
{A, B}	0.152	0.184	0.126	0.183
{B, C}	0.095	0.240	0.084	0.226
{A, C}	0.088	0.248	0.099	0.210

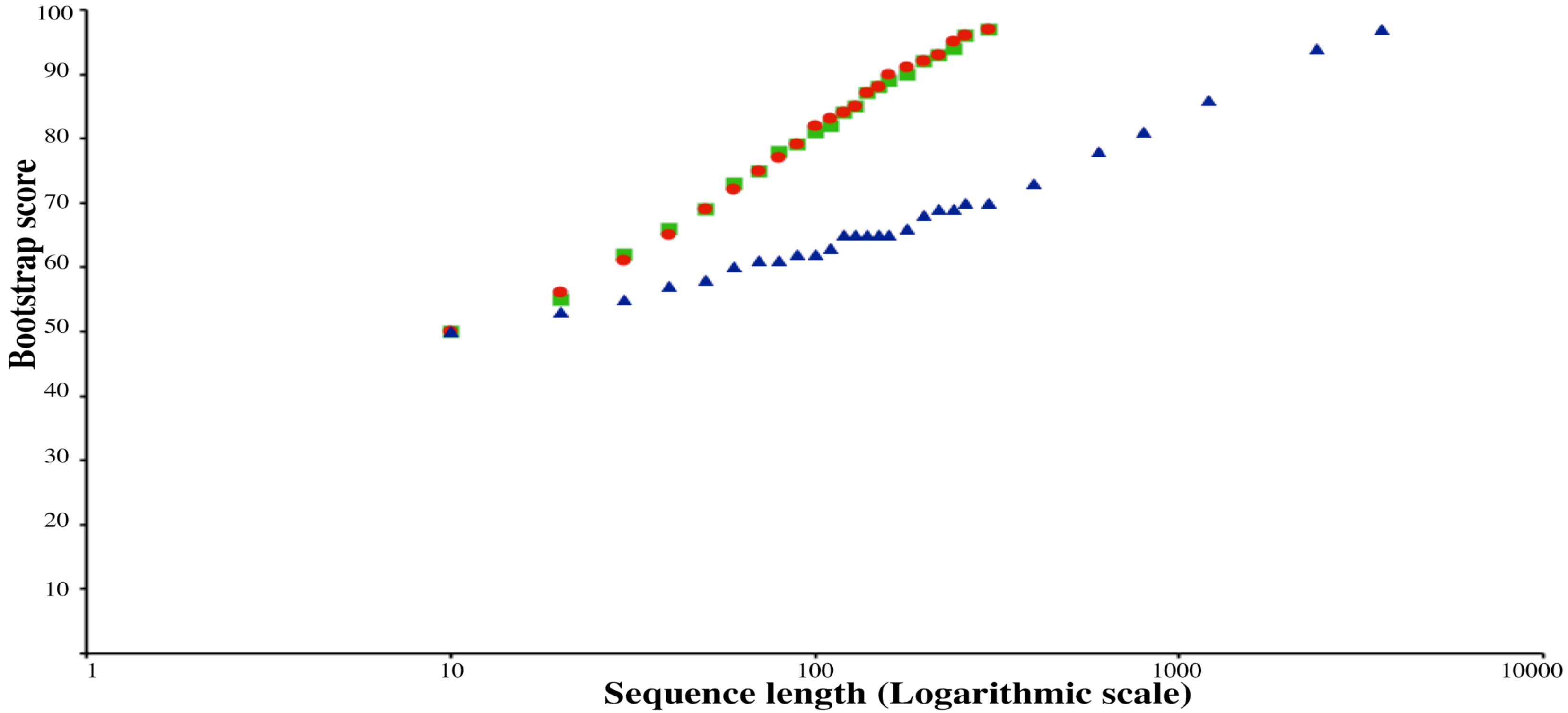


Figure 4.4: Graph illustrating effect of increasing sequence size (plotted along a logarithmic scale) on bootstrap supports for three different methodologies. Green squares and red circles represent supports from NJ and parsimony bootstrap while blue triangles represent supports from ML analyses. NJ and parsimony bootstrap converge on 100% rapidly while the length of the sequence must be significantly increased before maximum likelihood approaches 100%

4.4 Discussion

The *Neisseria* have been proposed as a model organism for researching transformation as they have been shown to readily uptake exogenous DNA (Koomey 1998). This ability to integrate foreign DNA into their own genome raises serious concerns for health researchers as it allows benign strains to acquire virulence factors that enable it to cause disease in its only host humans. It appears that individual *Neisseria* species are interconnected by a continuous horizontal flow of genetic information affecting chromosomal composition (Fussenegger *et al.* 1997). A secondary but still important implication of *Neisseria's* ability to transform is that it makes phylogenetic inference difficult, as it is nearly impossible to determine which genes are foreign and which are native. This study wished to investigate if there is any evidence for a sister-group relationship between any of the four completed *Neisseria* genomes. How best to analyse the data again raises some serious questions that have been debated in the literature for the last fifteen years. Broadly speaking there are two approaches that could be undertaken. The first involves a phylogenetic investigation of each individual gene while the second involves the concatenation of all data and then performing a phylogenetic analysis on this alignment in the belief that the real relationships will be found. A third (hybrid) approach is to investigate the data initially and then combine it if individual data partitions (genes) are congruent.

Taking the first approach, all single gene families that have phylogenetic signal (344 in total) according to two statistical tests were analysed. As each gene family contained four taxa, it follows that there are three possible rooted trees. Phylogenetic reconstruction of each gene leads to the finding that the frequencies of each topology were similar (36%, 33% and 31%). These frequencies are indicative of HGT, as there is a high degree of incongruence between individual gene trees. If HGT was absent and these bacteria were clonal (there is substantial evidence to shows this is not the case in *Neisseria* e.g. (Holmes *et al.* 1999)) then it should follow that one tree should be supported the majority of the time. Based solely on these observations, the third (hybrid) approach would not allow concatenation of these data, as it is very heterogeneous.

Even with the high levels of incongruence shown in the signal dataset, arguments from advocates of data concatenation would insist that the data should still be concatenated in the belief that the true underlying signal will be found (Barrett *et al.* 1991). Two datasets were concatenated together. The first concatenated alignment consisted of 344 genes ("signal" alignment), while the second consisted of the remaining 846 genes ("no signal" alignment) that were found to have no definite phylogenetic signal according to the relevant statistical tests. Historically the bootstrap procedure has been the most frequently used method for assessing the support for phylogenetic relationships (Cummings *et al.* 2003) therefore multiple bootstrap analyses of the "signal" concatenated alignment were performed and topology I (>80% BS) was inferred for ML, NJ and parsimony analyses with high BS. These results would lead one to believe that there is a single strong signal within the dataset as strong support for topology I is exhibited. The "no-signal" concatenated alignment of 846 genes supported topology III with 100% BS for both ML and NJ bootstrap analyses, parsimony bootstrap supported this grouping with 88% BS. There seems to be strong support for the grouping of *N. meningitidis* serogroups A and B (topology III) as sister species. To summarise the "signal" concatenated alignment infers a sister relationship for serogroups A and C while the "no signal" concatenated alignment infers a sister relationship for serogroups A and B. Both of these inferences are supported with high BS. The number of informative characters supporting each of the three possible topologies was determined using a retention index (Clyde and Fisher 1997). A near equal distribution of supporting characters among the three topologies is found for the "signal" alignment with a slight majority of informative sites (36%) supporting topology I. Similarly for the "no-signal" concatenated alignment the distribution of supporting characters among the three topologies was nearly equal except there is a slight majority (36%) supporting topology III. The high BS found for these datasets is a concern and must be the result of systematic errors associated with bootstrap analysis. It is not surprising that the topology with the majority of supporting characters is inferred but the bootstrap method appears to be too liberal in assigning high branch supports in this instance. Ideally phylogenetic methods should be efficient powerful robust and falsifiable, in terms of support this means that methods that need a minimum number of data to attain high supports values for true clades are preferable (Erixon *et al.* 2003). Examining the BS values and the frequency of informative characters from the concatenated data can only lead to the assumption that there are serious systematic problems associated with the BS resampling technique and it may not be as conservative as some have suggested (Zharkikh and Li 1992).

The results reported above show an inability of the Bootstrap analysis to assess all of the underlying phylogenetic information appropriately. Bootstrapping operates in a binary manner, i.e. it will either support a tree or not. This binary manner is the reason why when one topology has a small majority of supporting characters it can receive high bootstrap support. Resampling the data many times is supposed to overcome this shortfall, however the relationship between bootstrap support values and actual support seems to be more complicated. In an effort to visualise the possible pitfalls of a bootstrap analyses, computer simulations were preformed. The first simulation involved evolving a number of sequences (4 in all) of varying lengths along a tree. Each simulated sequence had three equal length partitions, each supporting one of the three possible topologies. Under these conditions, we would expect that no single topology should be favoured with high BS (>80%). This was not the case however as we observed approximately 25% false positive results (Table 4.1). When a false positive was observed it was generally the result of a tree having a slight majority of supporting characters (Table 4.2). It has been suggested that without some assessment of reliability a phylogeny has limited value (Sanderson 1995). The phylogeny may represent an efficient summary of available information based on character state distributions among taxa but it is uninformative regarding the evolutionary history of the taxa it encompasses (Sanderson 1995). The most widely used assessment of phylogeny reliability is the bootstrap and some have suggested that bootstrapping is a minimal requirement for any phylogenetic study (Penny and O'Kelly 1991). Based on the abnormalities observed within the concatenated data presented here, it was decided to perform a spectral analysis. A spectral analysis tests multiple hypotheses by quantifying both support for and conflict against definitive signals in the data (Lento *et al.* 1995). This approach is attractive as the sequence data is carefully examined for strongly supported alternative hypotheses instead of merely examining an optimal phylogenetic tree. Spectral analysis also provides qualitative and quantitative information on how strongly supported each hypothesis is contradicted. The spectral analysis performed here provides an accurate picture of true phylogenetic signal instead of noise generated through systematic errors obviously associated with the bootstrap technique employed. While I agree with that bootstrapping is a minimal requirement of a phylogenetic study, I also suggest that the underlying data should be examined carefully to overcome any systematic errors associated with this method.

The final simulation that was performed examined the effect that increasing amounts of data have on BS values. Initially a small alignment with three partitions was created. As the alignment length was increased, the bootstrap score increased (the relative frequency of supporting characters remained constant). This is not surprising and is reassuring to an extent. The parsimony and NJ bootstrap analyses quickly converged to 100% BS while the ML method did not converge as quickly. The ML analyses converged on 100% BS after the sequence length was increased above 3,000 nucleotides (Figure 4.4). The simulation studies provide evidence for positively misleading results by bootstrap analyses under certain conditions, in reality the tree chosen is the best representative topology but the support values associated with this tree seem to be inappropriately excessively. The results of this simulation agree with the findings of Bull *et al* (1993) who showed that as the number of inconsistent characters are added to a dataset the ability to recover the correct phylogeny decreases (Figure 4.5). These finding would seem to disagree with the idea that the real phylogeny will be recovered as more data are added, this scenario may only be true when consistent data are added (Figure 4.5). This is most likely an artefact of the bootstrapping methodology as topologies with slightly higher supports will be inferred more often and will therefore receive higher scores in the majority consensus tree. The inconsistencies demonstrated in this simulation are the result of systematic biases that have shown to be exacerbated when longer sequences are analysed (Phillips *et al.* 2004). Similar conclusions based on similar simulations have been reached by other authors (Bull *et al.* 1993). The significance of these simulations

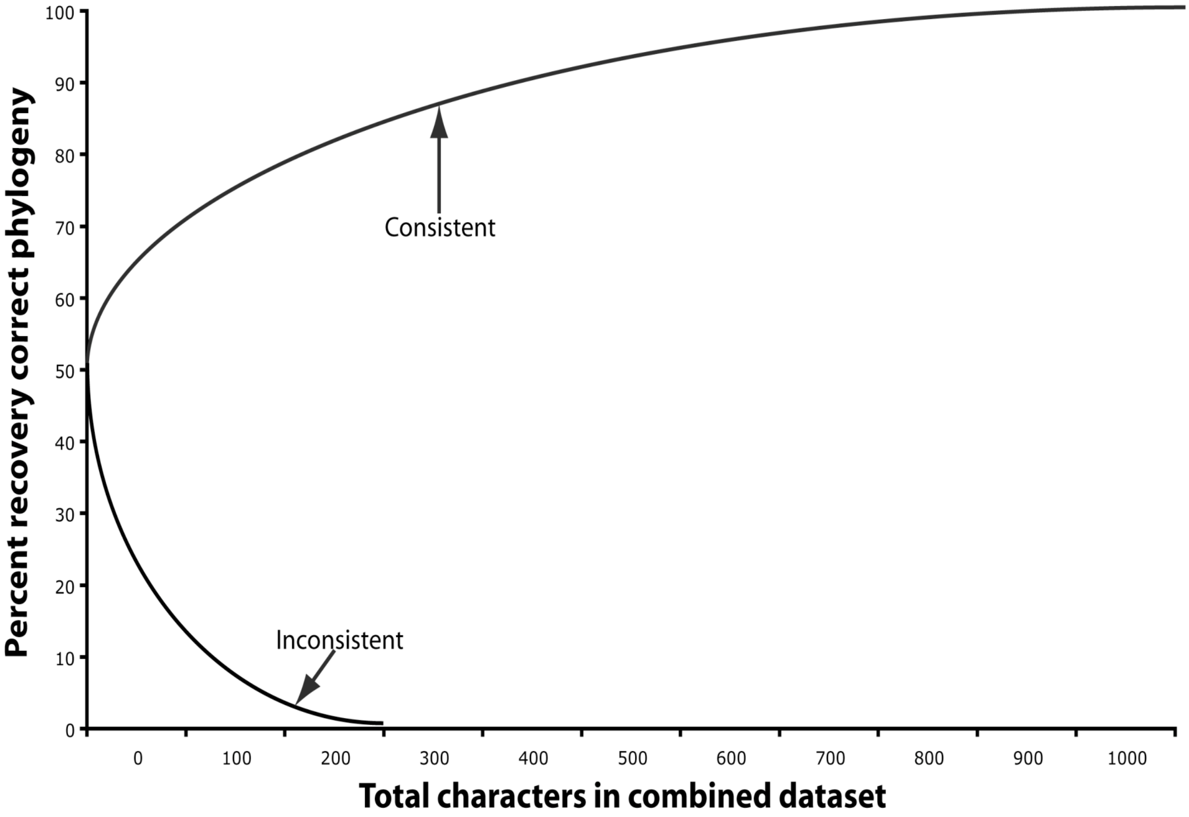


Figure 4.5: The effect of combining consistent data and inconsistent data. The probability of estimating the correct phylogeny is shown as a function of the number of characters in each dataset. Redrawn from Bull *et al* (1993).

has been questioned by Wiens (1998) where he states, “these simulations may have limited relevance to phylogenetic problems encountered in the real world. Most systematists typically examine more than four taxa for a given phylogenetic analysis”. While Wiens is correct in saying that the majority of phylogenetic studies incorporate more than four taxa, I do not think it is reason enough to ignore the anomalies shown by such simulations.

In conclusion, the above analyses illustrate the methodological problems that phylogenetisits face today when they wish to use a large number of genes to infer relationships. CC and TC give very different answers for the *Neisseria* dataset presented here. Which answer is correct? The CC analyses would lead us to believe that the *Neisseria* are clonal as particular groupings receive very high BS. There is a body of published and anecdotal evidence to strongly reject this hypothesis and the simulated studies carried out point to misleading conclusions a bootstrap analysis can provide (Suzuki *et al.* 2002; Erixon *et al.* 2003). Furthermore, as more data is added to an alignment the misleading conclusion will be supported with greater confidence. Alternatively looking at the individual gene trees, which have been shown to have real phylogenetic signal we see a huge degree of incongruence as all possible trees are found in near equal proportions. These high levels of incongruence are signatures of HGT. This conclusion is in agreement with a large body of published evidence that has shown the *Neisseria* undergo high levels of HGT. Opponents of CC will agree that this level of incongruence would have gone unnoticed if all data had been concatenated and is the reason data should never be combined. This leads onto the final method for analysing large datasets that has been proposed i.e. data concatenation after careful examination. Obviously, the levels of incongruence would not allow all the data to be combined. Therefore, I support the view that individual genes should be analysed before concatenation and genes should never be combined when there is evidence of HGT or high levels of heterogeneity. As an aside it is interesting to speculate as to what, these high levels of HGT mean for *Neisseria* virulence. As already mentioned vaccines have been developed for 4 of the 5 pathogenic *Neisseria* strains. It is entirely plausible that *Neisseria* strains may incorporate foreign genes, which are unaffected by current vaccines into their genomes thus reducing the current success attained through vaccination. To date there is no evidence to suggest that this is the case but perhaps vaccine designers should target essential house keeping genes, which are less prone to HGT.

Chapter 5 α-proteobacterial molecular phylogeny using a supertree approach and a phylogeny for mitochondria among the α-proteobacteria

5.1 Introduction

Proteobacteria or the purple bacteria are one of the largest known divisions within prokaryotes. Based on information from 16S rRNA sequences the proteobacterial division was first circumscribed in detail by Carl Woese and co-workers (Woese *et al.* 1985; Woese 1987). Subsequent 16S rRNA analysis has led to the proteobacteria being divided into five subdivisions (α, β, δ, γ, ε) (Woese 1987).

The proteobacterial group is an important division as it includes known animal, human and plant pathogens. Furthermore, this division has played a crucial role in eukaryotic cell origin and development through endosymbiosis. The hypothesis pertaining to an endosymbiotic origin of the mitochondria was first proposed in the19th century (Altman 1890) and there is strong evidence in contemporary literature to support the belief that present day mitochondria were once free-living α-proteobacteria (Andersson *et al.* 1998; Lang *et al.* 1999; Ogata *et al.* 2001), however the exact position of the mitochondria within the α-proteobacteria is still debated (Wu *et al.* 2004). A number of analyses have placed the mitochondria among the order Rickettsiales (Gupta 1995; Lang *et al.* 1999) while others propose that mitochondria are more closely related to the Rickettsiaceae family and *Rickettsia prowazekii* in particular to the exclusion of the Ehrlichia and Anaplasma genera (Karlin and Brocchieri 2000; Emelyanov 2003), which includes species such as *Wolbachia*. On completion of the *Wolbachia* genome sequence Wu and co-workers (Wu *et al.* 2004) reported strong support for a grouping of *Wolbachia* and *Rickettsia* as a sister group to the exclusion of the mitochondria. A recent analysis by Esser *et al* (2004) has led to the suggestion that current methods for constructing phylogenetic trees are insufficiently sensitive to ascertain the sister of mitochondria among the present sample of α-proteobacterial genomes. They concluded however that "*Rhodospirillum rubrum* comes as close to mitochondria as any α-proteobacterium investigated". Their analysis used 55 individual and 31 concatenated protein data sets encoded in *Reclimonas americana* and *Marchantia polymorpha* mitochondrial genomes.

Any attempt to determine which extant α-proteobacterium is the sister group of the mitochondria depends on the hypothesis that there is a robust and meaningful α-proteobacterial phylogeny, this in turn implies that within this group, horizontal gene transfer (HGT) has not been so frequent that a meaningful species phylogeny can be derived. Up to now single-gene phylogenies (especially SSU rRNA-based) have established many of the accepted relationships between bacteria. Single-gene analyses are dependent on that gene having an evolutionary history that reflects that of the entire organism. A better approach would be to combine information from many genes and to this end supertrees would seem to be a logical answer as they aim to combine information from many individual genes. The primary reason for the prerequisite of a robust supertree is to account for possible HGT events. HGT among bacteria has been shown to be a major source of genetic variation and the level of reported HGT incidences is continually increasing (Martin *et al.* 1998; Wolf *et al.* 1999; de la Cruz and Davies 2000; Brown 2003; Kinsella *et al.* 2003). If HGT is the dominant form of bacterial genome evolution it would prove fruitless to proceed with trying to identify the sister group of the mitochondria, as definitive relationships may be hard to prove. By investigating whether there is a single underlying phylogeny, we can gauge how often HGT may occur. If we know how much error we might expect, then we would know how much confidence we might invest in our conclusions regarding the mitochondrial relationships. Recently, evidence of congruent signal from multiple genes in closely related bacterial divisions (Daubin and Gouy 2001; Daubin *et al.* 2003) suggests that HGT in general has not been so severe, that it can wipe out phylogenetic signal among closely related species. However, early prokaryotic evolution cannot be represented effectively with a single organism phylogeny (Creevey *et al* 2004). Following this line of thought I have constructed an α-proteobacterial supertree derived from orthologous genes from complete genomes using the most similar supertree analysis (MSSA) method (Creevey *et* *al* 2004) (see Section 1.6.1 for a description of the method). This novel method was also used to assess confidence in the proposed supertree. Following supertree construction and estimation of the potential amount of confusion that could be created by HGT events, I then sought to identify the closest known relative of the mitochondrion endosymbiont.

5.2 Materials and Methods

5.2.1 Sequence data

The bacterial database used in this analysis consisted of seven completely sequenced α-proteobacterial genomes i.e. *Sinorhizobium meliloti*, *Agrobacterium tumefaciens*, *Wolbachia*, *Rickettsia prowazekii*, *Brucella melitensis*, *Caulobacter crescentus*, *Mesorhizobium loti* and the partial sequence data of *Novosphingobium aromaticivorans*, *Rhodobacter sphaeroides* and *Rhodospirillum rubrum*. *Magnetococcus magnetospirillium* homologues were used in three incidences in place of *R. rubrum* (explained in section 5.2.6). Sequence data for the partial genomes was supplied by DOE Joint Genome Institute (http://www.jgi.doe.gov/).

The mitochondrial database consisted of the above bacterial database as well as the complete mitochondrial genomes of *Reclinomonas americana* and *Marchantia polymorpha* as well as two outgroups in *Neisseria meningitidis* MC58 and *Escherichia coli* 0157:H7.

5.2.2 Database searches and alignments

For supertree construction, homologous sequences were identified by performing an all-against-all search of the bacterial database using the BLASTP algorithm (Altschul *et al.* 1997) with a cut off of 10^{-7}. From the multiple BLASTP searches, only those homologous single-gene families where every member found every other member (and nothing else) and had at least 4 taxa were retained. This conservative approach has been designed to minimise the inadvertent analysis of paralogs or spliced genes. In total, 537 single gene families were found. This dataset will be referred to as the supertree dataset from herein.

The set of proteins common to both *R. americana* and *M. polymorpha* were compared to the mitochondrial database using BLASTP (Altschul *et al.* 1997) with a cut off expectation value of $10^{-7.}$ This set of proteins will be referred to as the mitochondrial dataset herein. In some instances more than one match per genome was found. In these cases, the best match to the mitochondrial query was retained. This approach allows each α-proteobacteria to be as similar to mitochondria as possible at the level of sequence similarity and follows the procedure of Esser *et al* (2004). Homologues of atp9 were only detected after the BLASTP low complexity filter was switched off.

For each dataset, the individual protein families were aligned using ClustalW 1.81 (Thompson *et al.* 1994) using the default settings. For the mitochondrial datasets, gaps created in the amino acid alignments were inserted into the nucleotide sequences to produce codon based nucleotide alignments using the program putgaps (http://bioinf.may.ie/software/putgaps). All alignments were corrected for obvious alignment ambiguity using the alignment editor Se-Al 2.0a11 (http://evolve.zoo.ox.ac.uk/software.html?id=seal).

5.2.3 Gene tree construction

The phylogenetic relationships of the single gene families within the supertree dataset were reconstructed using a quartet puzzling approach and the WAG substitution matrix (Whelan and Goldman 2001) as implemented in Tree-Puzzle 5.1 (Schmidt *et al.* 2002). To ensure that there was phylogenetic signal within the single gene families a permutation tail probability (PTP) (Archie 1989) test as implemented in PAUP* 4.0b10 (Swofford 1998) was performed. Those genes that do not contain phylogenetic signal were eliminated from this study. This resulted in the supertree dataset being reduced from 507 gene trees to 406.

Phylogenetic relationships for the 31 mitochondrial-encoded genes (mitochondrial dataset) were reconstructed using a variety of methods. Bayesian trees derived from amino acid and codon based nucleotide alignments were constructed using MR BAYES 3.0B4 (Huelsenbeck and Ronquist 2001) with the among site rate variation set to invariable gamma. Clade probabilities for each phylogeny were determined using the *sumt* command of MRBAYES 3.0B4. ML phylogenies derived from both amino acid and codon-based nucleotides alignments were reconstructed using the WAG (Whelan and Goldman 2001) and HKY (Hasegawa *et al.* 1985) models of substitution as implemented in Tree-Puzzle 5.1. Branch supports were determined using puzzleboot (www.tree-puzzle.de/puzzleboot.sh), which determines the ML distance matrices of appropriately resampled data partitions. Additionally, NJ trees based on LogDet nucleotide distances (Lockhart *et al.* 1994) were reconstructed using PAUP* 4.0b10, with constant and third codon positions removed from all alignments. The LogDet distance was used in an attempt to account for compositional biases within sequences that may bring these biased sequences together in a clade even though they may not be each others closest relative. The LogDet distance is a measure based on the determinant of a divergence matrix (contains the relative frequencies of all characters) between two sequences. Initially it was thought that proteins were free from compositional biases, therefore LogDet distances have been restricted to the analysis of DNA sequences. Foster *et al* (1999) have shown that protein sequences can also contain biases leading to misleading results. For completeness phylogenetic hypotheses based on protein LogDet distances were constructed, this was a fruitless task because in many cases the amino acid frequency matrix results in one row or column being a linear combination of another, therefore it was impossible to calculate the LogDet distance. This problem arises when sequences do not contain enough information for such an analysis.

5.2.4 Supertree reconstruction

The best supertree was found by an exhaustive search of treespace using the MSSA method (Creevey *et al* 2004) as implemented in CLANN 2.0.2 (Creevey and McInerney 2004). To test the null hypothesis that the phylogenetic signal in the gene trees was better than random the YAPTP randomisation method (Creevey *et al* 2004) was utilised. The YAPTP method proceeds by replacing each gene tree with a randomly chosen topology with the same leaf set. This removes any congruent phylogenetic signal between gene trees, while leaving the numbers, sizes of gene trees, the frequency with which any particular taxon was found across the gene trees, and the frequency of co-occurrence of any group of taxa within gene trees unaltered (Creevey *et al* 2004). A heuristic search of tree space was then performed with the score of the best supertree being recorded. This was repeated 100 times. The null hypothesis that the gene trees contain no more phylogenetic signal than expected by chance alone is rejected if the score for the raw data is not bettered by any of the 100 sets of randomly permuted gene trees. In order to assess the support for internal branches on the supertree, a bootstrap analysis was performed. From the individual input trees a pseudo-dataset with the same number of input trees as the original dataset was created by resampling with replacement of the original dataset. For each pseudoreplicate a heuristic search of tree space was carried out and the results of this bootstrap analysis were summarised in a majority rule consensus tree. If all input trees are congruent with one another they should all be compatible with a single supertree (Creevey *et al* 2004). In order to examine the behaviour of such perfect data, fully compatible input trees (ideal dataset) were generated. For each input tree a corresponding ideal tree (one that fitted perfectly onto the supertree) was produced. The ideal trees replicate the taxonomic composition, frequency of co-occurrence and extent of overlap in the original input trees. An exhaustive search of supertree space was carried out using the ideal data and the scores of all possible supertrees were calculated. The reasoning behind this approach is that we can compare the distribution of tree scores obtained from real and ideal data and determine how close the data is to ideal (Creevey *et al* 2004).

5.2.5 Shimodaira-Hasegawa tests

In order to assess the likelihood that any differences in topology between the supertree and individual gene trees are no more significant that what are expected by chance Shimodaira-Hasegawa (SH) tests (Shimodaira and Hasegawa 1999) were carried out using Tree-Puzzle 5.1 (Schmidt *et al.* 2002). Gene trees varied in size from 4 to 10 taxa therefore when there were fewer than 10 sequences in any input dataset the supertree was appropriately pruned so that it contained the taxa that were found in the individual gene trees. The underlying amino acid alignment from which the input tree was derived was used for each SH test.

5.2.6 Concatenated mitochondrial dataset

Two concatenated mitochondrial encoded gene datasets were created. The first concatenated alignment followed the procedure of Esser *et al* (2004), *cox1*, *cox2* and *cox3* genes could not be located for *R. rubrum* therefore the appropriate homologues from *Magnetospirillum magnetotacticum* were used instead as *Rhodospirillum* and *Magnetospirillum* are almost always well-supported sister taxa (Esser *et al.* 2004). In total 31 gene families were found to contain all 14 taxa (*N. meningitidis* and *E. coli* outgroups, two mitochondria and ten α-proteobacteria). The complete list of genes is identical to those found by Esser *et al* (2004). The 31 gene families were concatenated (31-gene alignment) to yield an alignment of 11,184 amino acid sites or 33,552 codon based aligned nucleotides with a large number of gapped positions. Removal of all gapped positions produced a 6,724 amino acid alignment and according to Tree puzzle 5.1 (Schmidt *et al.* 2002) there is severe amino acid compositional heterogeneity across the sequences as all but two taxa (*M. loti and A. tumaficiens*) failed the χ^2 square test for homogenous amino acid composition at P=0.95.

The second concatenated alignment consists of the mitochondrially encoded genes whose phylogenetic relationships have a topology that is not significantly different to the proposed α-proteobacterial supertree according to a SH test. The reasoning behind this strategy is that these are the cohort of genes that do not appear to have any evidence of HGT and are probably the most reliable set of genes with which to examine the sister-group relationship of the mitochondrion. This second concatenated alignment consisted of 16 genes (Table 5.1) and in total has 5,050 aligned amino acid positions. Removal of gapped positions reduced the alignment to 3,311 amino acids. As with the 31 gene

Table 5.1: LogDet, ML and Bayesian results for 31 mitochondrial-encoded proteins. Gene names underlined and bold refer to the genes that have a topology that is not significantly different from the proposed α-proteobacterial supertree.

gene	sister of mitochondria	BP	sister of mitochondria	BP	sister of mitochondria	prob	sites
atp1	(ric,wol)	56	non-mono	-	(ric,wol)	1	533
atp6	(ric,wol)	44	(ric,wol)	100	(ric,wol)	1	321
atp9	(ric,wol)	82	(ric,wol)	100	(ric,wol)	1	79
cox1	(ric,wol)	55	(ric,wol)	100	(ric,wol)	1	704
cox2	(ric,wol)	57	ric*	49	ric*	0.87	460
cox3	(ric,wol)	50	(ric,wol)	100	non-mono	-	395
cob	(ric,wol)	93	ric	79	ric	0.97	494
nad2	(ric,wol)	93	ric	66	(ric,wol)	1	516
nad4	(ric,wol)	52	(ric,wol)	100	(ric,wol)	1	565
nad5	(ric,wol)	95	(ric,wol)	100	(ric,wol)	1	840
rpl2	(ric,wol)	86	ric*	89	ric	1	541
rpl5	(ric,wol)	51	(ric,wol)	98	(ric,wol)	1	236
rps14	(ric,wol)	55	(ric,wol)	100	(ric,wol)	1	122
yejU	(ric,wol)	97	ric	99	ric	1	327
rps11	(ric,wol)	85	non-mono	-	(ric,wol)	1	133
nad3	ric	33	ric	82	ric	0.94	187
rps2	ric	68	ric	100	ric	1	337
rps3	ric	27	ric	96	ric	0.76	477
rps8	ric	23	non-mono	-	wol	0.77	160
rps12	ric	60	ric	80	ric	0.79	148
rpl6	wol	59	wol	53	wol	1	205
rps1	wol	52	ric	74	ric	0.80	625
rps13	wol	62	wol	65	wol	0.75	135
rps19	wol	70	wol	61	wol	0.93	103
yejR	(ric,wol,neis)	17	non-mono	-	ric	1	859
nad4l	(ric,wol,og)	34	ric	84	wol	0.92	111
nad1	rru	82	ric*	96	ric	1	398
rps4	og	54	ric	92	ric	0.93	255
nad9	neis	79	ric	78	(ric,wol)	1	604
rpl16	non-mono[a]	-	ric	30	non-mono[b]	-	146
rps7	non-mono	-	ric*	60	(ric,wol)	0.51	237

Notes and abbreviations for Table 5.1

[a] *R. americana* branches with *R. prowazekii* and *Wolbachia*. *M. polymarpha* branches with the ougroups.

[b],*R .americana* and *M .polymorpha* are not monophyletic butbranch with *R. prowazeki* and *Wolbachi* which are are monophyletic

* The optimum topology according to a SH test.

ric : *R. prowazeki*, wol : *Wolbachia*, rru : *R. rubrum*

neis : *N. meningitidis*, og : *Escherichia coli* + *Neisseria meningitidis*

BP values indicate bootstrap supports

prob values indicate Bayesian clade probabilities concatenated this alignment fails the χ^2 square test for homogenous amino acid composition at P=0.95. The largest possible concatenated alignments in which all sequences passed the χ^2 test for homogeneity of amino acid composition were created. This was achieved using two methods that strip highly variable sites from the concatenated alignments. The first method described by Hansmann and Martin (2000) scores sites depending on the number of different amino acids that are found at that site. This effectively assumes that there is a star phylogeny with equal branch lengths. Sites with high scores were iteratively removed. Finally a homogenous alignment of 2,619 and 1,154 amino acids were constructed for the 31 and 16-gene alignments respectively. The second method is implemented in Tree-Puzzle 5.1 (Schmidt *et al*. 2002) and assumes 8 categories of sites with different rates of evolution and these rates are described according to a discrete gamma distribution (Yang 1994). Again fast evolving categories of sites were iteratively removed until alignments of 2,446 and 1,634 amino acid positions that passed the χ^2 test for homogeneity of amino acid composition were constructed for the 31 and 16-gene datasets respectively. Corresponding nucleotide alignments were also created for the 31 and 16 gene datasets that had had fast evolving sites removed (as determined using the discrete gamma distribution method). This resulted in three different types of compositionally homogenous alignments for each of the concatenated families: two protein alignments and a nucleotide alignment.

LogDet protein distances were determined for the homogenous alignments using LDDist (Thollesson 2004), the fraction of invariable sites was estimated by the method of Sidow *et al* (1992) and these were excluded. Neighbor joining (Saitou and Nei 1987) phylogenetic hypotheses based on these LogDet protein distances were constructed. One hundred bootstrap resamplings were also performed on the homogenous data to evaluate support for individual internal branches. Phylogenetic reconstructions of the nucleotide alignments were also created based on LogDet distances implemented in PAUP* 4.0b10 (Swofford 1998), invariant and third codon positions were removed. Phylogenetic hypotheses for all compositionally homogeneous alignments were also generated using a ML and Bayesian framework implemented in PAUP* 4.0b10 (Swofford 1998) and MrBayes 3.0B4 (Huelsenbeck and Ronquist 2001) respectively. As described earlier support values for internal branches were also determined. Phylogenetic networks were also inferred using the Neighbor-net method (Bryant and Moulton 2004). Phylogenetic networks permit the representaion of conflicting signal or alternative phylogenetic histories (Fitch 1997). When the evolutionary history of genes is not treelike it is necessary to use phylogenetic networks (Bryant and Moulton 2004). Even when the underlying history is treelike, parrell evolution and sampling error may make it difficult to determine a unique tree. In these cases networks provide a valuable tool for representing ambiguity or for visualizing a space of feasible trees (Bryant and Moulton 2004)

5.3 Results

5.3.1 α -proteobacterial supertree

The dataset in this analysis consisted of 10 α-proteobacterial genomes or 40,319 individual coding sequences. We identified 406 single gene families that met the criteria set out in section 5.2. Amino acid sequences were aligned resulting in a combined length of 126,057 aligned positions. Phylogenetic analyses for all 406 gene families were performed yielding individual ML gene trees. Using these 406 trees as input data, an exhaustive search of all possible 2,027,025 unrooted supertrees uniting all ten taxa was undertaken for both real and idealised data. For the real data, the range of supertree scores varies from 253 to 646 (Figure 5.1). The distribution of the scores of the 100 best trees from the YAPTP test is centred on 442(±28). No single tree from the YAPTP test receives a better score than the best supertree found from the exhaustive search. The signal from the input trees is better than random. Therefore, we assume that there is a meaningful α-proteobacterial phylogeny as the trees from the YAPTP test are significantly worse than the proposed supertree. The ideal data generated in this study represents a scenario where there is complete compatibility between the input trees and supertree. Looking at the distribution of tree scores for the ideal and real data in Figure 5.1, it is obvious that the real data has a high degree of congruence across the input trees

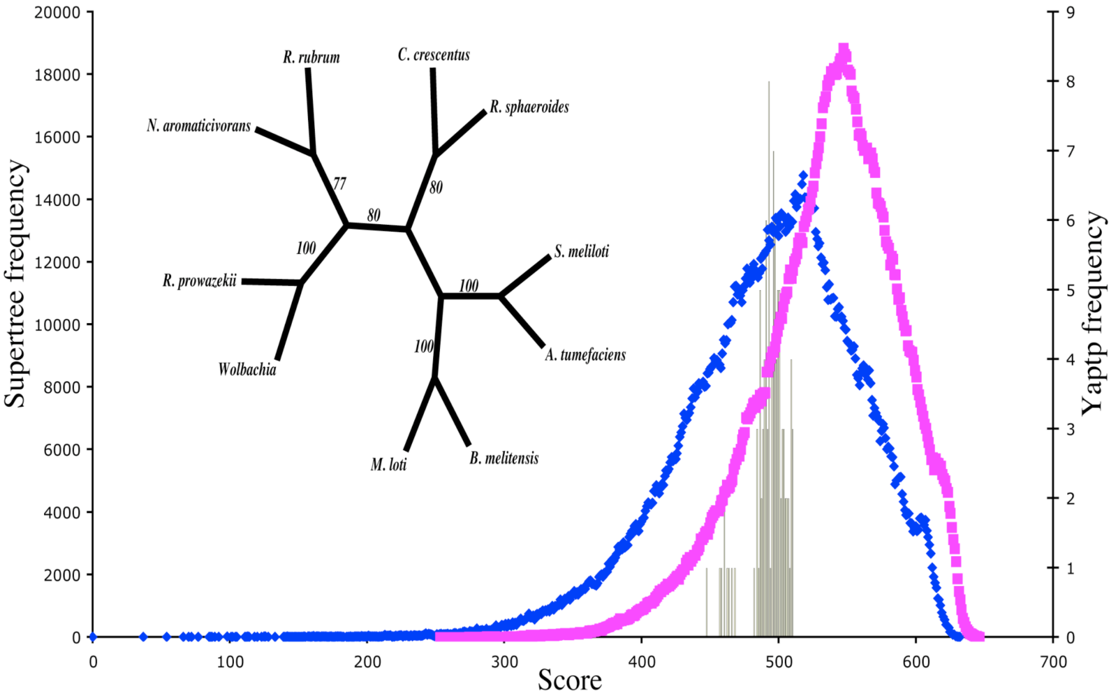


Figure 5.1: Analysis of 10 representatives of the α-proteobacteria. Pink dots represent the distribution of the supertree similarity scores of all possible unrooted trees. Blue dots represent the distribution for the idealised dataset. The tree within the figure is the supertree that achieved the best score for the real data; numbers on internal branches represent bootstrap support values. The histogram represents the distribution of the best similarity scores found for 100 repetitions of the randomisation test.

as it mirrors the distribution followed by the ideal data. The best supertree generated from the real data was found to have a score of 253 with only 0.001% of the idealised gene trees receiving a better score. The analysis described above was also performed using neighbor joining input trees created using PHYLIP (Felsenstein 1989), the results from these trees concur with the above results.

From the 406 input trees SH tests revealed that 105 (26%) described their underlying alignments significantly better than did the supertree as the differences in topology could not be explained by chance alone. From the remaining 301 trees, there is no significant difference between the proposed supertree and underlying tree (21% of the these are identical to the proposed supertree). I hypothesised that the 105 input trees that are significantly different from the supertree may be the result of HGT events. In an effort to quantify the types of genes that display significantly different topologies to the supertree I categorised genes as being either operational or informational in function based on the criteria used by Jain *et al* (1999). Following the complexity hypothesis (Jain *et al.* 1999) we would expect to see operational genes being exchanged at a higher rate than informational genes. Informational genes are members of large complex systems thereby making horizontal transfer of these gene products less probable (Jain *et al.* 1999). This hypothesis regarding the non promiscuous nature of informational genes has led researchers to create a robust bacterial phylogeny consistent with ribosomal RNA trees using informational genes exclusively (Brochier *et al.* 2002). In total our dataset was found to contain 84 informational genes and 94 operational genes, the remaining genes have yet to be assigned a function (i.e. conserved hypothetical) or belong to a category that is neither operational or informational. In total 16 of the 84 informational genes (20%) were found to have a topology that is significantly different to the proposed supertree, while 47 of the 94 operational gene (50%) trees differ significantly to the supertree according to the SH test. Assuming that there are no systematic or stochastic biases in phylogenetic reconstruction of the input trees and also assuming that the supertree is truly representative of the relationships among the α-proteobacteria, then this would indicate that the operational genes within this dataset experience significantly more horizontal transfers than informational genes.

From the input data, 104 alignments had all ten taxa, these alignments were concatenated to yield a 31,520 amino acid alignment. We performed 100 bootstrap resamples of this alignment based on maximum likelihood distances and protein LogDet distances with the removal of invariant sites. This resulted in a topology very similar to the best supertree found by the MSSA method with high branch support values (Figure 5.2) the obvious difference being that *N. aromaticovorans* and *R. rubrum* are not monophyletic and neither are *C. crescentus* and *R. sphaeroides*. The number of branch swaps needed to make these trees identical is therefore only two, a very small number considering that there are more than two million possible unrooted trees.

5.3.2 Mitochondrial origin

5.3.2.1 Individual mitochondrial encoded genes

Phylogenetic analysis of the 31 mitochondrial encoded genes were constructed for the codon based nucleotide using LogDet, ML and Bayesian methods. All three methods revealed a real consensus regarding the possible sister group of mitochondria (Table 5.1). The majority of these 31 gene trees placed either one or both of the bacteria from the Rickettsiaceae family as the sister group of the mitochondria. To assess which phylogenetic method most accurately represents the data we performed a SH test using the underlying alignment and the three proposed topologies for each gene family. An additional "constrained" tree where *R. rubrum* was forced to branch as a sister group to the mitochondria (recreated using MrBayes) was also assessed by the SH test, the reasoning behind this approach is that *R. rubrum* has been suggested as a possible sister group to the mitochondria (Esser *et al.* 2004). According to the SH test the "constrained" tree was significantly worse than the 3 alternative competing topologies in all cases and was therefore not considered for further analysis. The trees based on LogDet distances are considered the most reliable unless one of the alternative methods reproduced a

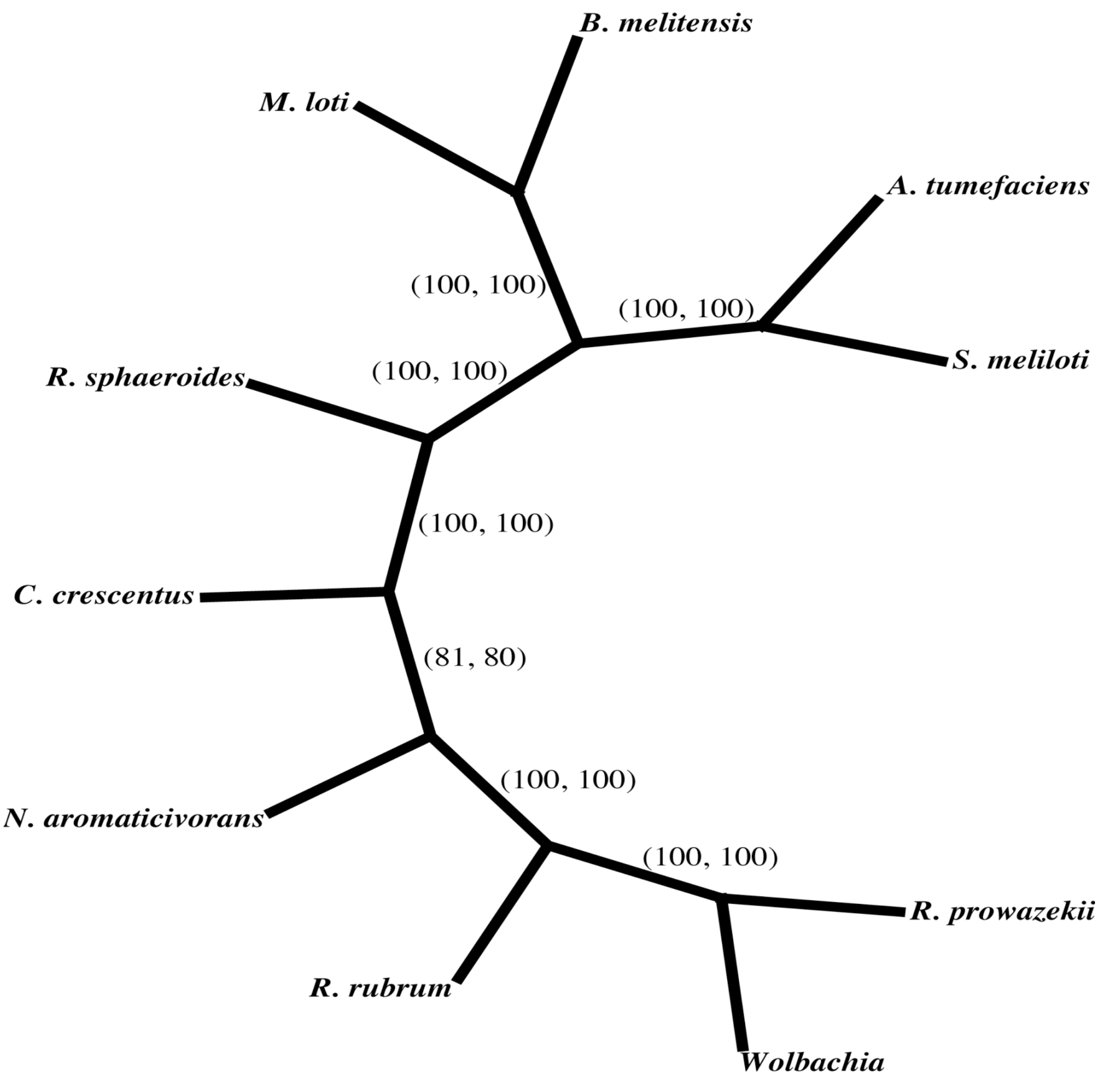


Figure 5.2: Phylogenetic tree inferred from 104 concatenated genes, each of which contains all ten α-proteobacteria taxa found within the bacterial database. Bootstrap supports are shown for internal branches, these were calculated based on LogDet and ML distances of the protein alignment.

phylogeny that fits the underlying data significantly better according to the SH test. In total 13 trees placed the group (*R. prowazekii* and *Wolbachia*) as the sister to the mitochondria. A total of 9 trees placed *R. prowazekii* as the sister group and 4 trees placed *Wolbachia* as the mitochondrial sister, 2 trees inferred a sister relationship with *R. prowazekii* and *Wolbachia* but also included the outgroups, from the remaining trees, 2 placed the outgroups as the mitochondrial sister. The final tree was not monophyletic for the mitochondria sequences (Table 5.1). To summarise 26 trees indicate that either the (*R. prowazekii* and *Wolbachia*) grouping or one of these bacteria exclusively is the closest extant relative to the mitochondria. Only a single LogDet gene tree (*nad1*) suggested that *R. rubrum* was the sister to the mitochondria and in this case another method was found to produce a significantly better tree where *R. prowazekii* was shown to be more closely related to the mitochondria than *R. rubrum.* All three phylogenetic methods reproduced in general similar trees for all 31 genes. The analysis described above was also performed on the protein alignments but I found it difficult to determine LogDet distances for the majority of these. Amino acid trees inferred by ML and Bayesian reconstruction criterion (Table 5.2) agreed in general with their nucleotide counterparts.

5.3.2.2 31-gene concatenated alignment

Concatenation of the complete mitochondrial dataset resulted in an 11,184 amino acid alignment. According to the amino acid heterogeneity test there is severe amino acid bias within this dataset therefore the assumption of stationary base frequencies (Gu and Li 1996) is not observed. Reasons for the violation of this assumption are obvious; four of the taxa are endosymbiont and have probably not evolved with the same pattern of nucleotide substitution as their free-living cousins since they last shared a common ancestor. To account for these biases, the LogDet transformation (Lockhart *et al.* 1994) and a modified Tamura-Nei (Tamura and Kumar 2002) distance matrix were used to infer the relationships among the α-proteobacteria and the mitochondria as these methods have been shown to have the ability to deal with amino acid biases. LogDet distances derived

Table 5.2: Individual ML and Bayesian results for 31 mitochondrial-encoded proteins. Analyses were performed on protein alignments. LogDet distances could not be determined for the majority of genes and are therefore not presented. The findings of these protein results compare favourably to their nucleotide counterparts

gene	sister of	BP	sister of	prob	sites
	mitochondria		mitochondria		
atp1	**(ric,wol, og)**	**98**	**(ric,wol, og)**	**1**	**1599**
atp6	**(ric,wol)**	**76**	**(ric,wol)**	**0.92**	**963**
atp9	**(ric,wol)**	**85**	**(ric,wol)**	**1**	**237**
cox1	**(ric,wol, og)**	**100**	**(ric,wol)**	**0.97**	**2112**
cox2	**ric**	**49**	**ric**	**0.82**	**1380**
cox3	**(ric,wol)**	**55**	**(ric,wol,rru)**	**0.54**	**1185**
cob	**(ric,wol)**	**45**	**(ric,wol)**	**0.58**	**1482**
nad2	**(ric,wol)**	**73**	**wol**	**0.62**	**1548**
nad4	**(ric,wol,og)**	**53**	**(ric,wol,og)**	**1**	**1695**
nad5	**(ric,wol)**	**65**	**(ric,wol)**	**1**	**2520**
rpl2	**ric**	**61**	**ric**	**0.70**	**1623**
rpl5	**(ric,wol)**	**39**	**wol**	**0.53**	**708**
rps14	**(ric,wol,og)**	**35**	**(ric,wol,og,novo)**	**0.63**	**366**
yejU	**ric**	**51**	**ric**	**0.94**	**981**
rps11	**wol**	**65**	**wol**	**0.87**	**399**
nad3	**ric**	**60**	**(ric,ecoli)**	**0.80**	**561**
rps2	**ric**	**38**	**(ric,wol)**	**0.69**	**1011**
rps3	**(ric,wol)**	**46**	**(ric,wol)**	**0.71**	**1431**
rps8	**ric**	**63**	**ric**	**0.61**	**480**
rps12	**(ric,wol)**	**55**	**(ric,wol)**	**0.80**	**444**
rpl6	**(ric,wol)**	**79**	**wol**	**0.81**	**615**
rps1	**wol**	**45**	**wol**	**0.63**	**1875**
rps13	**wol**	**74**	**wol**	**0.52**	**405**
rps19	**wol**	**41**	**wol**	**0.89**	**309**
yejR	**(ric,wol)**	**62**	**(ric,wol)**	**0.84**	**2577**
nad4l	**(ric,og)**	**78**	**(ric,og)**	**0.58**	**333**
nad1	**rru**	**36**	**rru**	**1**	**1194**
rps4	**og**	**75**	**(ric,wol,og)**	**1**	**765**
nad9	**og**	**46**	**og**	**0.60**	**1812**
rpl16	**wol**	**65**	**non-mono**	**-**	**438**
rps7	**og**	**56**	**non-mono**	**-**	**711**

from the amino acid alignments placed *R. prowazekii* and *Wolbachia* as the sister group to the mitochondria with 100% bootstrap support (Figure 5.3). LogDet distances and modified Tamura-Nei (Tamura and Kumar 2002) distances derived from the nucleotide equivalent alignment also gave a topology with 100% support for the grouping of the mitochondria with the two members of the Rickettsiaceae family (Figure 5.3), performing the same analyses on the alignment minus positions that contain gaps (6,725 aligned positions) yields near identical supports. The NeighborNet of protein LogDet distances for the concatenated alignments shows good support for the monophyly of the mitochondria, the Rickettsiaceae group and the outgroup as well as other well supported relationships among the free living α-proteobacteria in agreement with Esser *et al* (2004) (Figure 5.4).

5.3.2.3 31-gene concatenated alignment with fast evolving sites removed by method of Hansmann and Martin

Removing the fast evolving sites using the method of Hansmann and Martin (2000) from the 31-gene concatenated alignment resulted in an alignment 2,619 amino acids in length. The NeighborNet of LogDet protein distances for this shortened alignment places *R. rubrum* as the closest α-proteobacterium to the mitochondria as it shares a split with *M. polymorpha* (Figure 5.5), this finding concurs with those found by Esser *et al* (2004). To assess the degree of support for these groupings, 100 bootstrap replicates were generated. Protein LogDet distances were determined for each replicate and phylogenies constructed using the NJ method and summarized using a majority-rule consensus tree. Again, we found results, which concur with those of Esser *et al* as *R. rubrum* is inferred to be the sister of the two mitochondrial genomes in 64/100 replicates. However when we used the ML and Bayesian reconstruction methods we found that the (*R. prowazekii* and *Wolbachia*) group to be the sister group to the mitochondria with 86% bootstrap support and 0.81 clade probability respectively, in this case there are obvious differences due to the tree reconstruction method used.

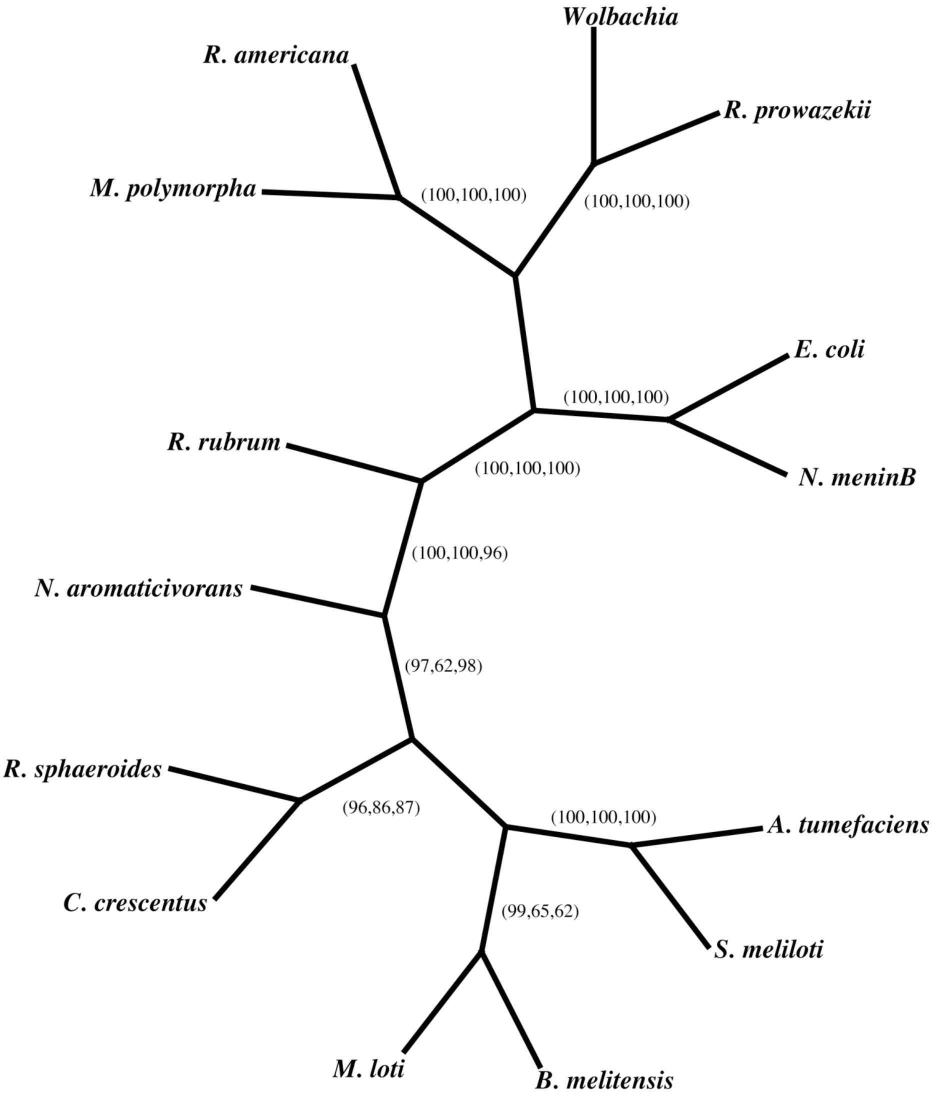


Figure 5.3: Phylogenetic tree inferred from a concatenated alignment of 31 proteins common to *R. americana* and *M. polymorpha* mitochondrial genomes and also present in ten α-proteobacteria and two outgroups. Supports on branches are bootstrap supports based on in order LogDet distances of amino acids and nucleotides and also modified Tamura and Nei distances (Tamura and Kumar 2002) of nucleotides.

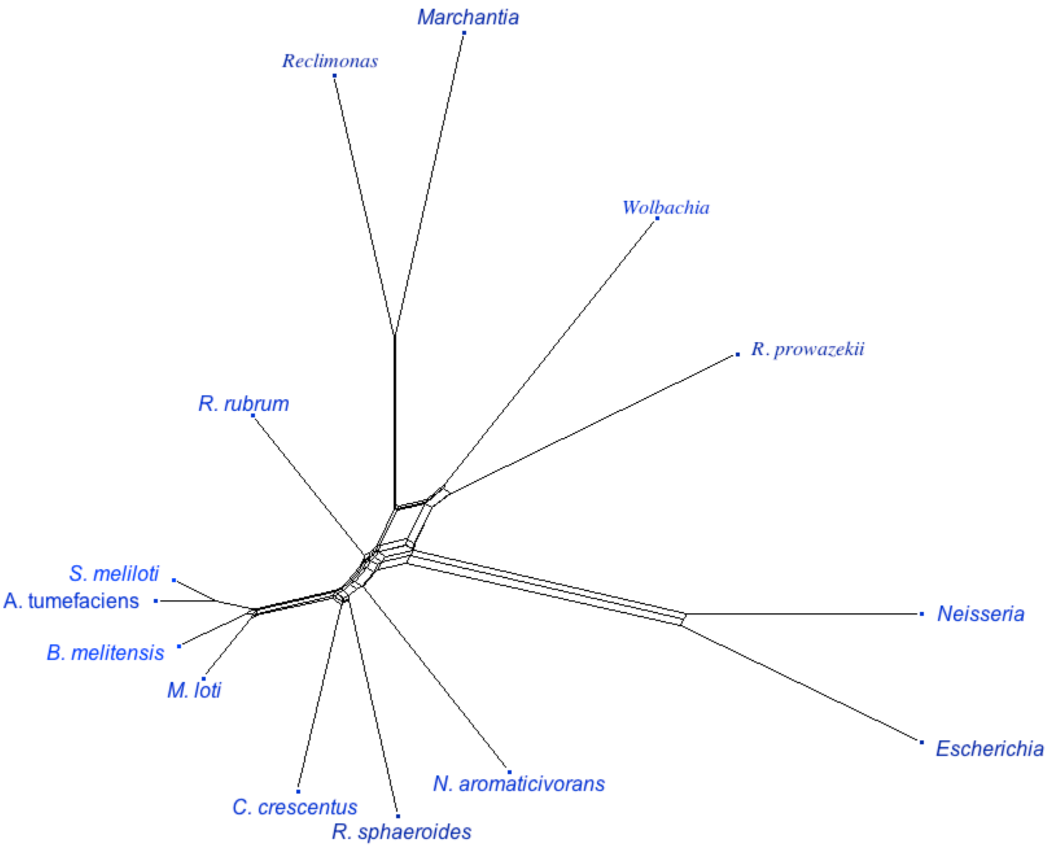


Figure 5.4: NeighborNet planar graph of 31 concatenated genes, Sites that contained gaps were removed from the alignment, this yielded 6725 aligned amino acid positions. Distances were determined using LogDet transformation with invariant sites removed. *Rhodospirillum* does not share a split with the mitochondria.

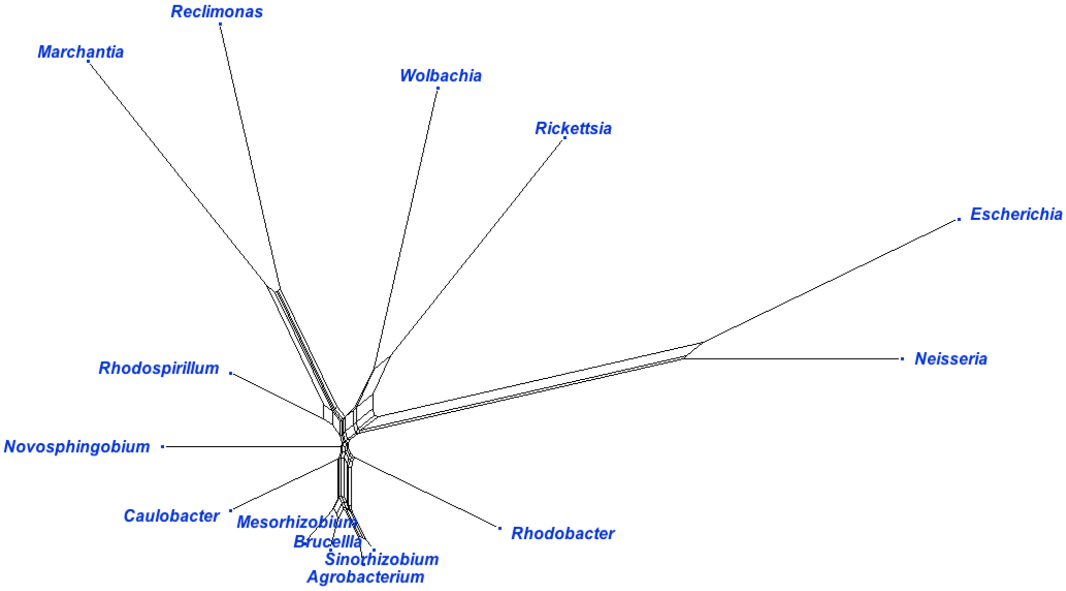


Figure 5.5: NeighborNet planar graph of 31 concatenated genes after fast evolving sites have been removed using the method of Hansmann and Martin (2000). Distances determined using LogDet transformation with invariant sites removed. *Rhodospirillum* shares a split with the mitochondria.

5.3.2.4 31-gene alignment with fast evolving sites drawn from a gamma distribution removed

Removing the fast evolving sites categorised using a discrete gamma distribution from the 31-gene concatenated alignment resulted in an alignment 2,446 amino acids in length. The NeighborNet of LogDet protein distances for this second shortened alignment conflicts with the inferred relationships from first shortened alignment (described in section 5.3.2.3) that has had sites removed using the method of Hansmann and Martin (2000), as it places *R. rubrum* closest to the Rickettsiaceae and it does not actually share a split with the mitochondria (Figure 5.6). The nucleotide equivalent of this shortened alignment was available therefore I calculated the nucleotide LogDet distance matrix (third positions and invariant positions removed). The NeighborNet representation of this matrix does not place *R. rubrum* beside the mitochondria but instead lends strong support to the grouping of the Rickettsiaceae and the mitochondria as they share a split (Figure 5.7).

The choice of phylogenetic reconstruction method did not effect the inferred topology as all three methods (LogDet NJ, ML and Bayesian) placed the *R. prowazekii* and *Wolbachia* grouping beside the mitochondria with relatively low support (43% and 77% BP for LogDet NJ and ML and 0.99 clade probability using the Bayesian framework). As a precaution, trees derived from nucleotide based codon alignments were also reconstructed. Again, the *R. prowazekii and Wolbachia* grouping was found to share its most recent common ancestor with the mitochondria in all cases except this time branch supports were unequivocal (100% BP for both LogDet NJ and ML and a clade probability of 1 found by the Bayesian approach). LogDet distances have a proven ability to deal with biased base frequencies, there is an obvious GC bias in the concatenated alignment used in this analysis (Table 5.3). Looking at figures 5.6 and 5.7 it is obvious that long branch attraction may be responsible for the grouping of the mitochondria with the Rickettsiaceae. A transversion analysis was also performed on the nucleotide alignment by removing third codon positions and recoding nucleotides as purines and

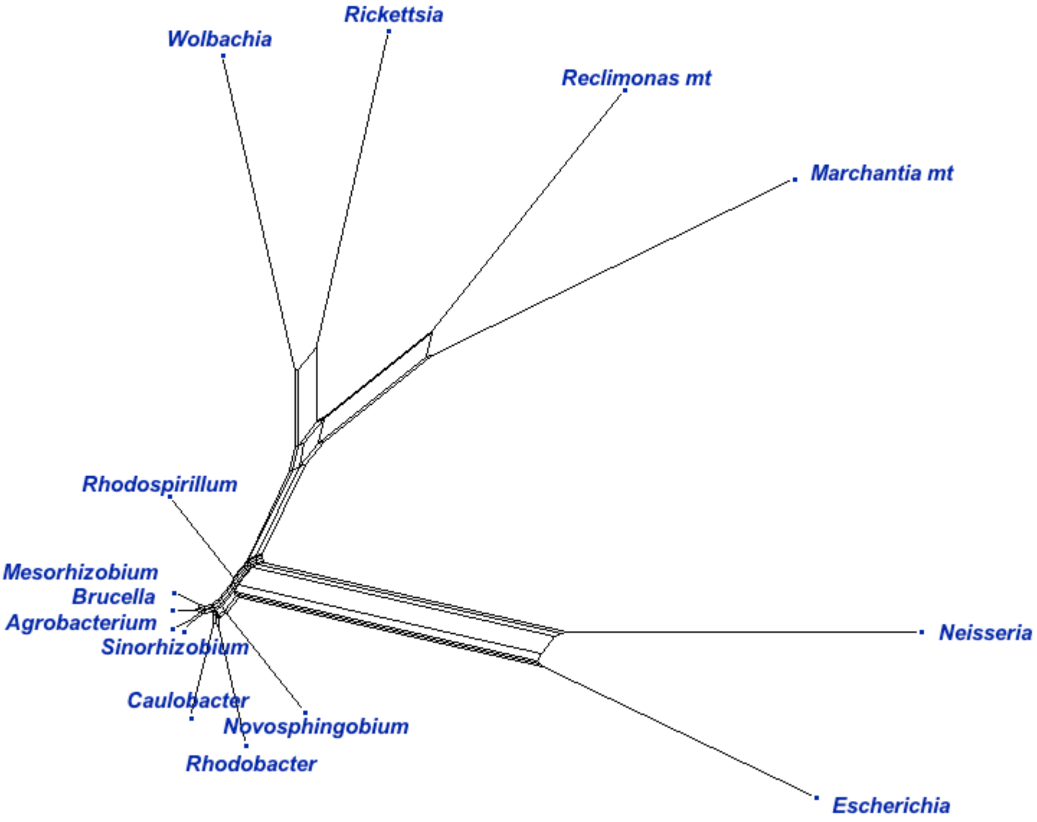


Figure 5.6: NeighborNet planar graph of 31 concatenated mitochondrial encoded genes (amino acid alignment). Fast evolving sites were categorised using a gamma distribution and iteratively removed until a homogenous alignment was found. Distances were determined using the LogDet transformation, invariant positions were excluded. The mitochondria and Rickettsiaceae are sister groups. *Rhodospirillum* does not share a split with the mitochondria.

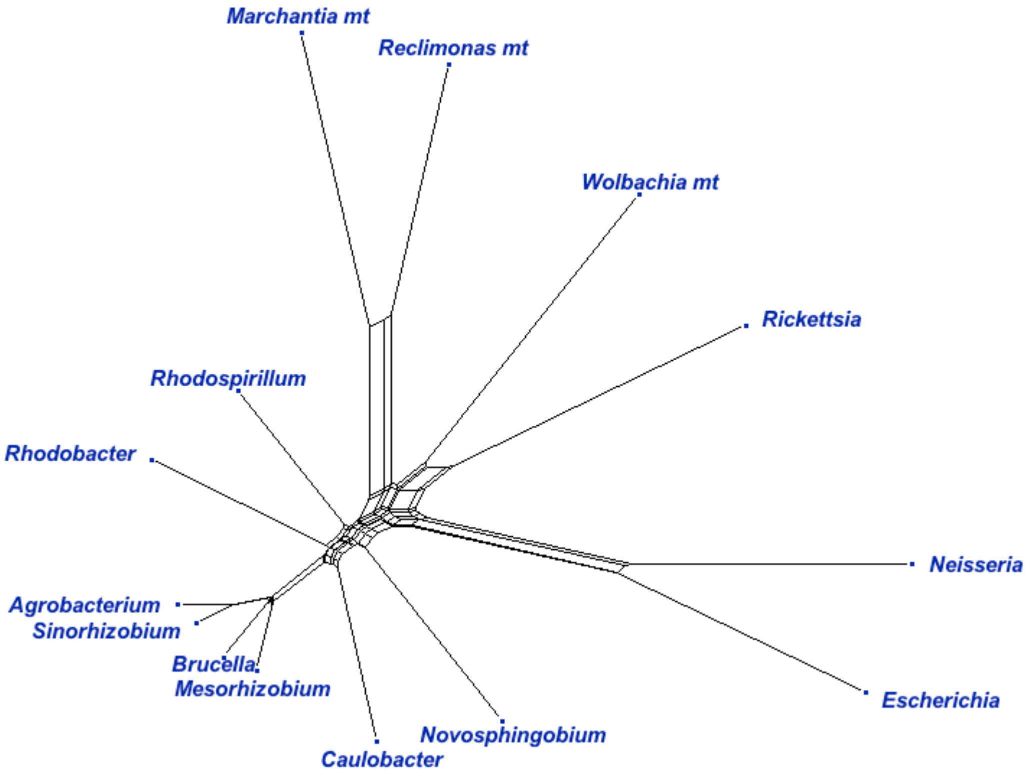


Figure 5.7: NeighborNet planar graph of 31 concatenated mitochondrial encoded genes (nucleotide alignment). Fast evolving sites were categorised using a gamma distribution and iteratively removed until a homogenous alignment was found. Distances were determined using the LogDet transformation, invariant positions were excluded. The mitochondria and Rickettsiaceae are sister groups. *Rhodospirillum* does not share a split with the mitochondria.

Table 5.3: Base frequencies of 14 taxa. Original alignment has had fast sites categorised by a discrete gamma distribution. These have been iteratively removed until a homogenous alignment which passes a chi square test of homogeneity was found. The mitochondrial and Rickettsiaceae sequences have a GC content that differs significantly to the other taxa represented.

Taxon	A	C	G	T
N. meningitides	0.20101	0.25838	0.27596	0.26465
E. coli	0.19338	0.27582	0.29150	0.23930
R. prowazekii	0.28455	0.16081	0.23072	0.32393
Wolbachia	0.26397	0.17253	0.24898	0.31453
R. americana	0.30431	0.14922	0.20673	0.33974
M. polymorpha	0.26751	0.18057	0.23331	0.31862
R. rubrum	0.16394	0.32829	0.30390	0.20387
C. crescentus	0.16190	0.34383	0.31153	0.18275
A. tumefaciens	0.16953	0.31916	0.28945	0.22186
S. meliloti	0.16789	0.33320	0.29817	0.20074
M. loti	0.16394	0.33742	0.30499	0.19365
N. aromatovarcians	0.16558	0.34206	0.30485	0.18752
R. spaeroides	0.16299	0.34819	0.31085	0.17798
B. meletensis	0.17171	0.30431	0.29013	0.23385
Mean	**0.20301**	**0.27527**	**0.27865**	**0.24307**

pyrimidines. This method reduces the branch lengths and gives high support for the grouping of the Rickettsiaceae with the mitochondria (Figure 5.8).

5.3.2.5 16 gene concatenated alignment

Of the 31 mitochondrial encoded genes 16 have a phylogeny that is either identical to the proposed supertree or the differences in topology are no more significant than expected by chance according to the SH test. Concatenation of these 16 genes gives a 5050 amino acid alignment. As with the 31-gene alignment there is severe amino acid bias according to the amino acid heterogeneity test. The LogDet transformation and the NJ method of phylogenetic reconstruction was used to infer the relationships among the α-proteobacteria and the mitochondria. LogDet distances derived from the amino acid alignment placed *R. prowazekii* and *Wolbachia* as the sister group to the mitochondria with 83% bootstrap support (Figure 5.9). LogDet distances (invariant and 3rd positions removed) derived from the nucleotide equivalent alignment groups the mitochondria with the two members of the Rickettsiaceae with 100% support (Figure 5.9). Performing the same analyses on the alignment minus positions that contain gaps (3311 aligned positions) yields near identical supports.

5.3.2.6 16-gene concatenated alignment with fast evolving sites removed by method of Hansmann and Martin

Removing the fast evolving sites using the method of Hansmann and Martin from the 16-gene concatenated alignment resulted in an alignment 1,154 amino acids in length. The NeighborNet of LogDet protein distances for this shortened alignment is bush-like; this may be due to the lack of phylogenetic information within this alignment although nearly 23% of sites are parsimony informative. To assess the degree of support for these groupings 100 bootstrap replicates were generated and the LogDet distance matrices used to infer NJ phylogenies were summarized as a majority rule consensus tree. This analysis showed that the Rickettsiaceae, *R. rubrum* and the mitochondria all share a clade with relatively low bootstrap support (59%).

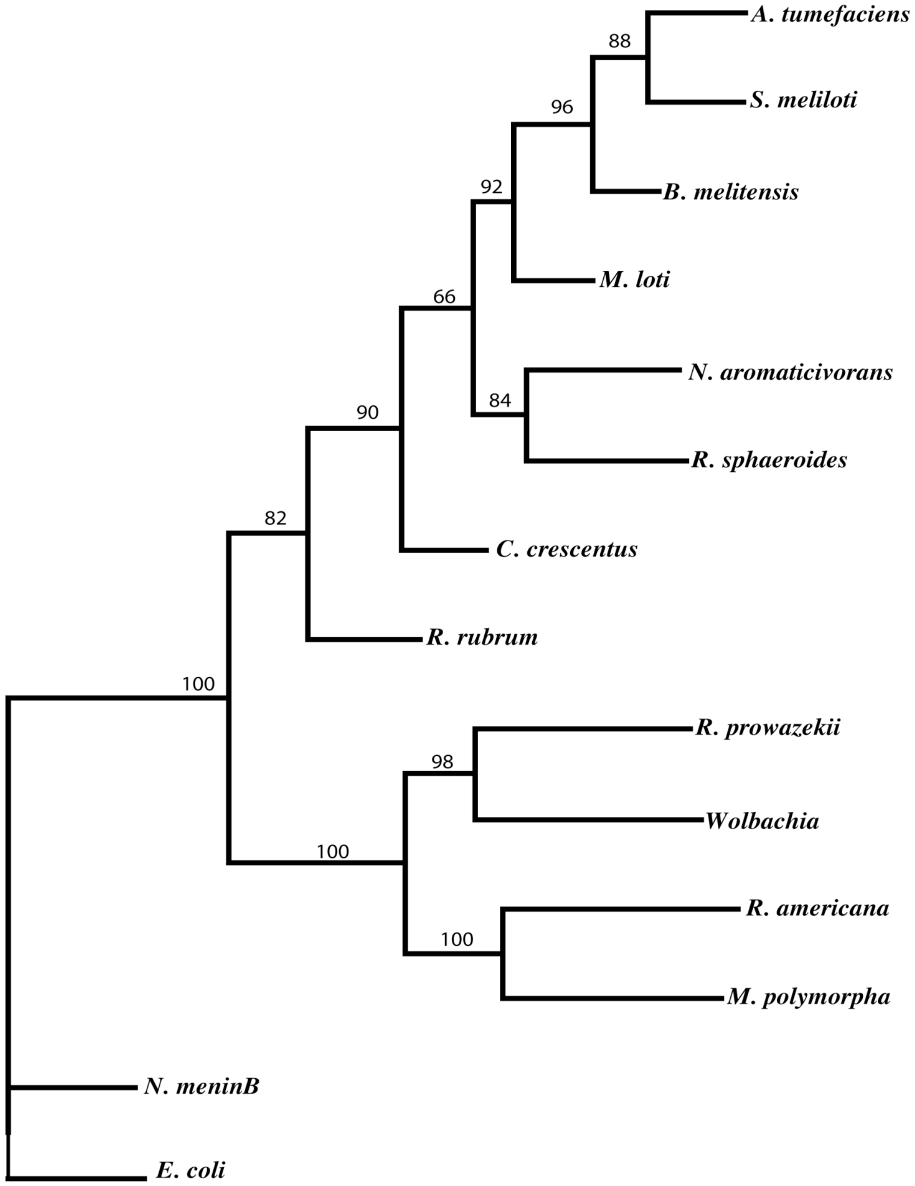


Figure 5.8: Maximum parsimony phylogenetic tree derived from the concatenated alignment of 31 mitochondrial encoded. Purines and pyrimidines have been recoded as R and Y. This type of analysis can account for possible GC bias. Note the relatively high support for the Mitochondria and Rickettsiaceae.

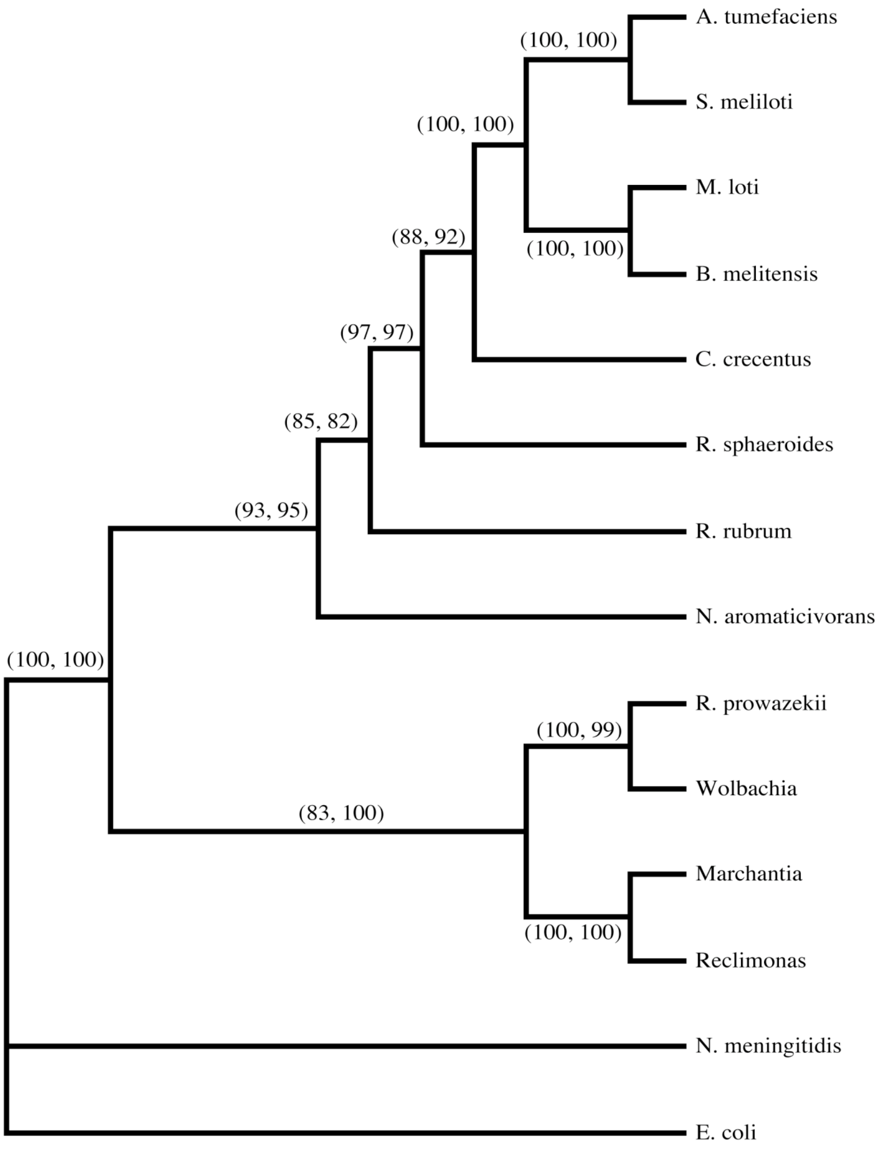


Figure 5.9: Phylogeny for 16 concatenated mitochondrial encoded proteins who all have a topology which is not significantly different (according to a SH test) to the α-proteobacterial supertree proposed earlier. Distances are based on LogDet distances. Numbers is brackets show branch supports calculated using nucleotide and amino acid concatenated alignments respectively.

5.3.2.7 16-gene alignment with fast evolving sites drawn from a gamma distribution removed

Removal of the fast evolving sites determined by a gamma distribution from the 16-gene concatenated alignment resulted in an alignment 1,634 amino acids in length. The NeighborNet of LogDet protein distances for this shortened alignment places the Rickettsiaceae beside the mitochondria as they share a split (Figure 5.10), this finding is similar to what was found for the 31-gene alignment that had fast evolving sites removed in this manner also. To assess the degree of support for these groupings 100 bootstrap replicates of this data were generated and the corresponding LogDet distance matrices were calculated. This analysis showed that the Rickettsiaceae and the mitochondria are each others closest relatives with 90% bootstrap support. As the nucleotide equivalent alignment was available I also performed a LogDet bootstrap analysis on this (invariant and 3rd positions were removed). The *R. prowazekii and Wolbachia* grouping was found to branch with the mitochondria again, this time with 100% bootstrap support. Tree topology is identical to Figure 5.9.

5.4 Discussion

In this analysis, a phylogeny for ten α-proteobacteria derived from 406 gene trees was proposed. This tree receives a better score (253) than any tree from the randomisation procedure (YAPTP) and the gene trees have substantially better fit to the optimal supertree. The score distributions from real and idealised gene trees are remarkably similar. We find that when we prune the optimal supertree to the size of each individual gene tree, 75% of them show no significant difference. The remaining 25% of trees that show a significant difference may be result of systematic biases in reconstructing phylogenetic relationships for individual gene trees or else horizontal gene transfer. I have shown that operational genes seem to be more promiscuous that informational genes, as a significantly higher percentage of operational genes disagree significantly

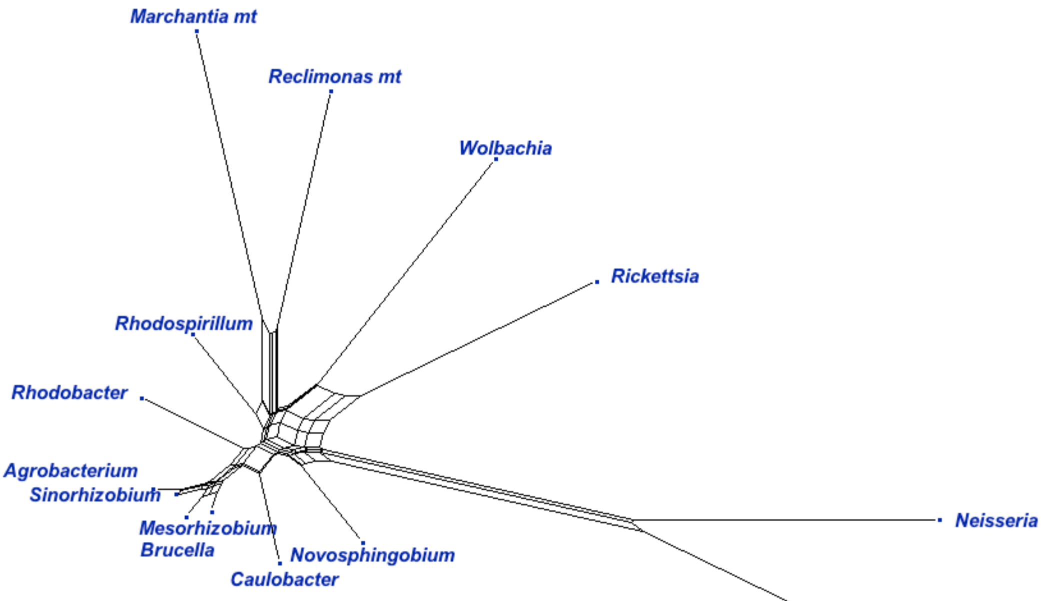


Figure 5.10: NeighborNet planar graph of 16 concatenated mitochondrial encoded proteins who all have a topology which is not significantly different (according to a SH test) to the α-proteobacterial supertree proposed earlier. Fast evolving were categorised using a discrete distribution and iteratively removed until a homogenous alignment was found. Distances were determined using the LogDet transformation, invariant positions were excluded. *Rhodospirillum* does not share a split with the mitochondria.

with the supertree, a finding in agreement with the complexity hypothesis (Jain *et al* 1999).

Recently *R. rubrum* has been suggested as the possible sister of the mitochondria based on a concatenated alignment of 31 mitochondrial-encoded genes (Esser *et al* 2004). Following the methods used by the above authors I found results in agreement with their findings. However using the same concatenated alignment and removing the fast evolving sites determined by a gamma distribution gives findings that disagree totally with those stated and instead points to a sister relationship between the Rickettsiaceae and the mitochondria with varying degrees of support, depending on if the amino acid or nucleotide alignments are used. These results raise concerns regarding the method used to strip out fast evolving sites. The method described by Hansmann and Martin (2000) simply strips sites out based on the characters found at a particular site while the discrete gamma method uses maximum likelihood estimation to place sites into different site categories. Clearly, these methods categorise sites very differently and the merits of each will need further study. Esser *et al* (2004) also made the point that the overall fermentative physiology of *Rhodospirillum* is quite similar to eukaryotes that lack mitochondria or contain anaerobic mitochondria and furthermore *Rhodospirillum* branched as the sister to the two mitochondria in 65/100 replicates. To address this observation I have shown that the Rickettsiaceae branch as the sister group when an alternative method to strip out variable sites is used. Furthermore, other authors have shown that there is a strong affinity between mitochondria and *Rickettsia* for the genes coding for components of the Krebs cycle (Andersson *et al.* 1998), the respiratory chain (Gray *et al.* 1999) and the translation system (Sicheritz-Ponten *et al.* 1998).

These results show that for the mitochondrial datasets there is evidence that the mitochondria originated from a direct ancestor of the Rickettsiaceae as a large majority of the 31 individual trees (83%) displays a sister relationship to the mitochondria for either both or one of the bacteria from this genus. Phylogenetic inferences based on highly conserved mitochondrial proteins such as cytochrome oxidase subunits (*cox1*, *cox2* and *cox3*) and apocytochrome b (*cob*) have been suggested as better determinates of relationships than fast evolving mitochondrial genes that may infer incorrect relationships as an artefact of long branch attraction (Lang *et al* 1999). Taking these four genes alone we can see that they all have a phylogeny that would suggest that the mitochondria and Rickettsiaceae are each others closest relatives (Table 5.1), the relative bootstrap supports for this relationship are low however (55%, 57%, 50% and 93% respectively).

A better understanding of the possible origin and evolution of the mitochondria will only be attained when a more representative sample of the α-proteobacteria are fully sequenced. It is entirely possible that there are many more mitochondrial like species yet to be identified. At present it is estimated that approximately 0.4% of all existing bacterial species have been identified and formally described let alone sequenced. However this analyses of individual gene trees and concatenated alignments leads to the conclusion that it is unparsimonious to suggest that the eukaryote mitochondria were once free-living ancestors of *R. rubrum*. Instead based on the evidence shown here and the available sequence data I conclude that instead the mitochondria most probably originated from an ancestor of a member of the modern day Rickettsiaceae.

Chapter 6 General Discussion

6.1 Comparative genomics

With the advent of whole genome sequencing a new revolution in molecular evolution research has begun. Determining the DNA sequence of a complete bacterial genome is now possible (depending on size and coverage) in less than two years (Weinstock 2000). With continued advances in automation and informatics this time frame as well as the associated costs should decrease. The impact large scale genomics has had on the landscape of modern biology cannot be understated. Genomics has forced the creation of novel technologies and methodologies to exploit this new wealth of sequence information. One such novel methodology is comparative genomics. Comparative genomics requires the input of multiple genomic sequences is revolutionizing our view of the microbial world and is providing insights into bacterial pathogenicity and evolution. This study aimed to utilise and develop computational and comparative methods to investigate mechanisms of evolution of a number of bacterial genera. I also wished to concentrate on specific genes and speculate about the mechanisms that help them evade host recognition. Furthermore, the bacterial ancestor of the eukaryote mitochondrion as well as the prevalence of horizontal gene transfer were addressed.

6.2 Large scale search for positive Darwinian selection

Immediately after the first data on protein sequences and electrophoretic mobilities became available in the 1960s evolutionists who wanted to understand and quantify Darwinian selection rushed to study these molecules. These selectionists were to be disappointed however, as they mostly observed purifying selection. The importance of purifying selection was first predicted by Schmalhausen (1949) and subsequent analysis on the new data led Kimura to publish his neutral theory of molecular evolution (Kimura 1968). The neutral theory of molecular evolution predicts that the majority of DNA variation within and between species is neutral with respect to fitness and can be described by stochastic fluctuations in a finite population (Kimura 1983). Furthermore, adaptive mutations can contribute to improvement of genes more directly than neutral mutations though they occur at an extremely low frequency (Endo *et al.* 1996). Most deleterious mutations on the other hand, do not contribute to gene evolution because they have a high chance of being lost.

In an essay predicting the future of biology, Haldane suggested that it should be possible to find evidence for the central mechanism of Darwinian evolution namely adaptive change (Sharp 1997). However apart from a few sporadic cases there has been little evidence that positive Darwinian selection is nature's *modus operandum* (i.e. there are ~100 papers reporting incidences of positive Darwinian selection). With the advent of better methods for detecting such selection and the increase in genomic data, it should be feasible to truly determine the relevance of Darwinian selection *in vivo*.

In Chapter 2, a large scale search for bacterial genes on which positive selection operates was performed. This large-scale search used the methodologies (distance based) of previous studies (Endo *et al.* 1996) as a reference point but also introduced new maximum likelihood methodologies that can account for heterogeneous selection among codons (Yang *et al.* 2000). An automated pipeline was developed to perform all the relevant tasks associated with such a search.

Initially it was hoped that all available prokaryote genomes could be combined to yield a large database and from this genes of interest could be found. This approach had two major pitfalls. Firstly, rates of evolution in homologous bacterial genes appear to be very rapid. Therefore, alignments were sometimes ambiguous and unreliable for the maximum likelihood methods used, as they assume the genes to be closely related and not to suffer from saturation of synonymous sites (Zhieng Yang; personal communication). Secondly, even when a reliable gene family was located it was, in some cases, too large for such an analysis as the time associated with performing the maximum likelihood calculations increases exponentially as the number of sequences are added. For these two reasons it was decided that each of the databases tested would be limited to four closely related bacterial genomes. This approach was not without its own complications as simulations in the past have shown that for the maximum likelihood methods to be very robust a minimum of six sequences need to be present before valid conclusions can be made (Anisimova *et al.* 2001). Simulations were performed to investigate these claims, the results of the simulations showed that the ML software was robust when only four sequences and strict criteria regarding the parameters indicative of positive selection were used therefore providing validation, in part, for the approach taken. The four genera examined were *Neisseria*, *Chlamydia*, *Bacillus* and *Escherichia*. Overall 4,324 single gene families were examined and from these 126 (~3%) were shown to have undergone an adaptive event according to the ML analysis. Conversely, the distance-based methods previously used by Endo *et al* (1996) only detected a single gene family as having undergone an adaptive event. The percentages of genes shown to have undergone positive selection in this analysis are similar to those found in chordate (5.3%) and embryophyta gene families (3.6%) (Liberles *et al.* 2001).

Before this analysis was performed, I expected that a significant number of the genes inferred to be under the influence of positive selection would be membrane associated. The reasoning for this hypothesis is that membrane-associated proteins are in direct contact with environmental and host pressures. While a number of membrane-associated proteins were found to have undergone positive selection the number (9 or ~7%) was lower than what I would have expected. A number of genes involved in the construction of pili from a number of genera were shown to have undergone positive selection. The pili are the only organelles that protrude from the bacterial membrane surface and are therefore in direct contact with the host immune system. The 75 genes from the *Neisseria* dataset that were shown to exhibit positive selection were selected for further study in an attempt to determine how many of these are due to differences in urogenitary and nasopharynx strains. Overall, 22 genes were inferred by both ML and a parsimony method (Creevey and McInerney 2002) as having undergone adaptive evolution due to speciation. The parsimony method was used as a precaution as the ML method that tests for selection along specific lineages has been shown to give a high number of false positives (Zhang 2004). For 11 of the 22 genes inferred by both methods, data was available regarding the biological pathways in which they are involved. In the absence of empirical and experimental data it is impossible to draw definitive conclusions as to the significance of these 22 genes, it was found however that 2 positively selected genes are involved in the biosynthesis of tryptophan and therefore this amino acid may be in limited supply in the urogenitary tract.

To summarise, an automated pipeline that utilises ML methods has been developed to test for positive Darwinian selection. The proportion of bacterial genes inferred to be evolving under the influence of positive Darwinian selection is similar to certain eukaryote groups. The gene families in this analysis were restricted to four sequences. If this analysis was performed on a larger set of genomes it is reasonable to assume that the observed proportion of positively selected genes would increase. Similarly, an increase in the number of taxa present in the study would allow for the use of alternative methods that can detect for positive selection (i.e. conservative parsimony based methods). The results from this analysis are in general agreement with what is predicted by the neutral theory, an increase in sequence data could alter these findings however.

6.3 Analysis of bacterial OMPs

One of the greatest achievements of humankind in the 20th century was the use of antimicrobial drugs to control infectious diseases. Penicillin and sulphonamides initially dominated the antibiotic era. Within two decades following the introduction of penicillin most of today's classes of antibiotics had been discovered (Loferer *et al.* 2000). Furthermore, no new chemical classes of active antibiotics have been successfully introduced into the clinic for over 30 years (Hancock 1998). Microorganisms have shown an amazing versatility in overcoming the affects of synthesised antibiotics and short-term measures such as chemical modification of existing antibiotics will have some impact on antibacterial therapy in the immediate future (Loferer *et al.* 2000). However, it is evident that new antimicrobial targets are required. To this end, large scale genomics has revolutionised the way research is conducted in this field. For example, following the sequencing of *N. meningitidis* serogroup B researchers located open reading frames that potentially encode surface-exposed or exported proteins (Pizza *et al.* 2000). From these candidate ORFs they successfully cloned and expressed a total of 350. Finally, proteins that elicit bactericidal activity were tested as candidate antigens for conferring protection against heterologous meningococcal strains. The approach taken by these researchers is commendable as it specifically targets proteins and should reduce the time and cost associated with developing antimicrobial agents. However, previous studies have shown that vaccines targeted against variable epitopes are less successful that those targeted against highly conserved epitopes (Suzuki and Gojobori 1999; Holmes *et al.* 2002). To this end I have tested for evidence of positive Darwinian selection (which can lead to variation) in 7 meningococcal vaccine targets as well as a highly conserved OMP that has been previously suggested as having the desirable characteristics of a vaccine target (Stephens and Lammel 2001; Genevrois *et al.* 2003; Voulhoux and Tommassen 2004).

As already mentioned OMPs are exposed to host immune responses, therefore it is reasonable to assume that they experience pressure to change. The first gene family examined consisted of ten pathogenic δ-proteobacterial Omp85 homologues. Omp85 is found in all gram negative bacteria and is essential for cell viability as it orchestrates protein placement within the outer membrane (Kleinschmidt 2003). Using a heterogeneous ML method (Yang *et al.* 2000) and a parsimony based sliding window (Fares *et al.* 2002) I determined that the essential transmembrane regions of the protein which anchor it in the outer membrane and under very strong purifying selection while a number of loops that cross the membrane experience selective pressures for change. These results illustrate the inability of essential regions of protein to change while regions, which may not be as important structurally are free to change, especially if it aids evasion of host defences.

Seven OMPs from the *Neisseria* genus were also examined for any evidence of positive Darwinian selection. These proteins are presently undergoing phase I clinical trials (Grandi 2003). The analysis performed showed five of the proteins were under very strong purifying selection while the remaining two exhibited evidence of having undergone positive selection. These findings have serious implications for the design of vaccines. If it transpires that newly developed vaccines target variable epitopes, then this could possibly lead to the vaccine becoming obsolete. Therefore, I suggest that a relatively inexpensive evolutionary study should be performed on all vaccine targets in the future. In this manner, evolutionists can help vaccinologists target highly conserved epitopes that cannot change due to structural constraints.

6.4 *Neisseria* phylogeny

Presently there are four completely sequenced genomes from the *Neisseria* genus. In Chapter 4 I posed the question is there a phylogeny among these four isolates? The *Neisseria* make up an interesting dataset as it is well known that they readily uptake foreign DNA via transformation (Koomey 1998). Therefore, due to horizontal gene transfer determining definite relationships is not a facile operation. Indeed when the 344 genes that have definite phylogenetic signal (according to two statistical tests) were examined it was found that for each of the three possible rooted topologies were supported approximately equally, an observation indicative of horizontal gene transfer. Presently there is an ongoing debate as how best to examine large datasets such as the *Neisseria* data described above. Supporters of an approach termed taxonomic congruence (Kluge 1989; Eernisse and Kluge 1993) advocate individual examination of each gene tree and a subsequent consensus analysis. In the case of the 344 *Neisseria* genes a consensus tree would be unresolved. An alternative approach termed character congruence (Kluge 1989; Eernisse and Kluge 1993) advocates the concatenation of data into a single alignment and then examination by an appropriate phylogenetic reconstruction method. Following this approach I found that one topology is supported with a very high bootstrap support value. Of the two approaches described, which best describes the *Neisseria* data? Taxonomic congruence indicates high levels of heterogeneity within the data, which is unsurprising considering the reported levels of transformation within this genus. Alternatively character congruence would seem to suggest that there is one strongly supported topology and that vertical and not lateral transmission of genes is prominent. The concept of vertical transmission of *Neisseria* genes would seem to disagree with a large body of published data (Maiden and Feavers 1995). Possible reasons why the total evidence approach would return a single topology needed to be addressed. Careful examination of all parsimony informative sites revealed that the proportion of sites supporting the 3 possible topologies were nearly equally distributed. A slight majority (36%) supported the topology that received high bootstrap scores. Based on these observations alone I was not surprised that this topology was inferred however I did not expect the bootstrap scores to be so high. These results raised concerns regarding stochastic errors associated with the bootstrap technique.

To address these concerns, simulations were performed and these unearthed serious methodological problems associated with the bootstrap technique. The results of this simulation were in general agreement with previous studies (Sanderson 1995; Erixon *et al.* 2003). Depending on which optimality criterion used (i.e. Parsimony, ML or NJ), support values converged to 100% BS rapidly as length of the sequence was increased. In my view, methodological problems such as these make concatenation of data uninformative and in the worst case they will infer the wrong result with strong support. If researchers insist on concatenating data they should carefully examine the underlying signal within the data with alternative methods such as spectral analyses (Lento *et al.* 1995). A spectral analysis determines the level of support and conflict for particular relationships and therefore does not operate in the simple binary "support or don't support" manner of the bootstrap.

6.5 α-proteobacterial supertree

The endosymbiont theory regarding the origin of the mitochondria is now widely accepted (Emelyanov 2001). This theory states that the mitochondria can trace its descent to a free-living bacterial ancestor that once entered into an endosymbiotic relationship with an ill defined, primitively amitochondriate cell (Emelyanov 2001). The invading bacteria has subsequently lost or passed the majority of its genes to the host. The end result of which is the modern mitochondrion, the main ATP-producing organelle of eukaryotes. Comparative studies of mitochondrial genomes unequivocally point to a eubacterial ancestry of mitochondria (Lang *et al.* 1999). Their monophyletic nature and close relationship to the Rickettsiales of α-proteobacteria emerged from phylogenetic reconstructions based on conserved proteins and SSU rRNA (Kurland and Andersson 2000). A phylogenetic analysis of chaperonin 60, likely the best tracer of the eubacterial origin of the mitochondria (Emelyanov 2001) which involved all Rickettsial sequences known has been published (Emelyanov and Sinitsyn 1999). This analysis demonstrated the paraphyletic nature of the Rickettsiales and the closest relationship of the genus *Rickettsia* to mitochondria (Emelyanov and Sinitsyn 1999). On completion of the *Wolbachia* genome sequence Wu and co-workers (Wu *et al.* 2004) reported strong support for a grouping of *Wolbachia* and *Rickettsia* as a sister group to the exclusion of the mitochondria, therefore the mitochondrion ancestor was most likely a member of the order Rickettsiales but not necessarily a Rickettsiaceae . A recent analysis by Esser *et al* (2004) has led to the suggestion that among the present sample of α-proteobacterial genomes "*Rhodospirillum rubrum* comes as close to mitochondria as any α-proteobacterium investigated".

In chapter 5, the data and methods used by Esser *et al* (2004) were investigated. Using their methodologies I recreated results in agreement with their findings. However, using the same data but removing fast evolving sites using a different methodology to that used by these authors yields conclusions that conflict with their findings. It is entirely possible that the mitochondion is actually descended from a family within the α-proteobacteria that has yet to be described and not from the Rickettsiaceae family as is widely thought. However based on the data used in this analysis this does not appear to be the case. Only increased genome sampling from the α-proteobacteria can help us answer this question definitively.

Any attempt to determine which extant α-proteobacterium is the sister group of the mitochondria depends on the hypothesis that there is a robust and meaningful α-proteobacterial phylogeny. If HGT is the dominant form of α-proteobacterial evolution it would prove fruitless to proceed with trying to identify the sister group of the mitochondria. To investigate this claim I constructed an α-proteobacterial supertree based on all available sequence data. By investigating whether there is a single underlying phylogeny it was possible to gauge how severe the effects of HGT are. Recently, evidence of congruent signal from multiple genes in closely related bacterial divisions (Daubin and Gouy 2001; Daubin *et al.* 2003) suggests that HGT in general has not been so severe, that it can wipe out phylogenetic signal among closely related species. However, early prokaryotic evolution cannot be represented effectively with a single organism phylogeny (Creevey *et al* 2004). I concluded that there is a robust α-proteobacterial phylogeny and from this I found that 16 of the 31 mitochondrial encoded proteins that do not have a topology significantly different from the proposed α-proteobacterial supertree. Concatenation of these genes and subsequent phylogenetic network and tree analysis inferred that *R. rubrum* is not the sister taxon to the extant mitochondria. Instead based on the available evidence the Rickettsiaceae family is the sister taxon to the mitochondria.

Chapter 7 Bibliography

Adams, E. N., Consensus techniques and the comparison of taxonomic trees. Systematic Zoology **21**: 390-397

Altman R. (1890). Die Elementarorganismen und Ihre Beziehungen Zur Den Zellen. Keipzig, Germany: Verlag von Veit.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research **25**(17): 3389-402.

Andersson, S. G., A. Zomorodipour, J. O. Andersson, T. Sicheritz-Ponten, U. C. Alsmark, R. M. Podowski, A. K. Naslund, A. S. Eriksson, H. H. Winkler and C. G. Kurland (1998). The genome sequence of Rickettsia prowazekii and the origin of mitochondria. Nature **396**(6707): 133-40.

Andrews, T. D. and T. Gojobori (2004). Strong positive selection and recombination drive the antigenic variation of the PilE protein of the human pathogen Neisseria meningitidis. Genetics **166**(1): 25-32.

Anisimova, M., J. P. Bielawski and Z. Yang (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Molecular Biology and Evolution **18**(8): 1585-92.

Anisimova, M., J. P. Bielawski and Z. Yang (2002). Accuracy and power of bayes prediction of amino acid sites under positive selection. Molecular Biology and Evolution **19**(6): 950-8.

Anisimova, M., R. Nielsen and Z. Yang (2003). Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics **164**(3): 1229-36.

Archie, J. W. (1989). A randomization test for phylogenetic information in systematic data. Systematic Zoology **38**: 251-278.

Arwert, F. and G. Venema (1973). Transformation in Bacillus subtilis. Fate of newly introduced transforming DNA. Molecular & General Genetics : **123**(2): 185-98.

Bainton, R., P. Gamas and N. L. Craig (1991). Tn7 transposition in vitro proceeds through an excised transposon intermediate generated by staggered breaks in DNA. Cell **65**(5): 805-16.

Bainton, R. J., K. M. Kubo, J. N. Feng and N. L. Craig (1993). Tn7 transposition: target DNA recognition is mediated by multiple Tn7-encoded proteins in a purified in vitro system. Cell **72**(6): 931-43.

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert and W. F. Doolittle (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. Science **290**(5493): 972-7.

Barany, F., M. E. Kahn and H. O. Smith (1983). Directional transport and integration of donor DNA in Haemophilus influenzae transformation. Proceedings of the National Academy of Sciences of the United States of America **80**(23): 7274-8.

Barrett, M., M. J. Donoghue and E. Sober (1991). Against consensus. Systematic Zoology **40**: 486-493.

Barthelemy, J. P., and F. R. McMorris, (1986) The median procedure for n-trees, Journal of Classification, 3(2):329-334.

Baum, B. R. (1992). Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon **41**: 3-10.

Bayer, M. E. (1968). Areas of adhesion between wall and membrane of Escherichia coli. Journal of General Microbiology **53**(3): 395-404.

Berezin, C., F. Glaser, J. Rosenberg, I. Paz, T. Pupko, P. Fariselli, R. Casadio and N. Ben-Tal (2004). ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics **20**(8): 1322-4.

Berg, D. E. and M. M. Howe (1989). Mobile DNA. American Society for Microbiology.

Bierne, N. and A. Eyre-Walker (2004). The genomic rate of adaptive amino acid substitution in Drosophila. Molecular Biology and Evolution **21**(7): 1350-60.

Bininda-Emonds, O. R. and H. N. Bryant (1998). Properties of matrix representation with parsimony analyses. Systematic Biology **47**(3): 497-508.

Bininda-Emonds, O. R., J. L. Gittleman and A. Purvis (1999). Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). Biological Reviews of the Cambridge Philosophical Society **74**(2): 143-75.

Bininda-Emonds, O.R.P., J.L. Gittleman, and M.A. Steel. (2002). The (super)tree of life: procedures, problems, and prospects. Annual Reviews of Ecology and Systematics 33: 265-289.

Bjune, G., E. A. Hoiby, J. K. Gronnesby, O. Arnesen, J. H. Fredriksen, A. Halstensen, E. Holten, A. K. Lindbak, H. Nokleby and E. Rosenqvist (1991). Effect of outer membrane vesicle vaccine against group B meningococcal disease in Norway. Lancet **338**(8775): 1093-6.

Boucher, Y., C. J. Douady, R. T. Papke, D. A. Walsh, M. E. Boudreau, C. L. Nesbo, R. J. Case and W. F. Doolittle (2003). Lateral gene transfer and the origins of prokaryotic groups. Annual Review of Genetics **37**: 283-328.

Brochier, C., E. Bapteste, D. Moreira and H. Philippe (2002). Eubacterial phylogeny based on translational apparatus proteins. Trends in Genetics **18**(1): 1-5.

Brochier, C., P. Forterre and S. Gribaldo (2004). Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the Methanopyrus kandleri paradox. Genome Biol. 2004; 5(3): R17.

Brochier, C., P. Lopez-Garcia and D. Moreira (2004). Horizontal gene transfer and archaeal origin of deoxyhypusine synthase homologous genes in bacteria. Gene **330**: 169-76.

Brown, J. R. (2003). Ancient horizontal gene transfer. Nature Review Genetics **4**(2): 121-32.

Brown, J. R. and W. F. Doolittle (1997). Archaea and the prokaryote-to-eukaryote transition. Microbiology and Molecular Biology Reviews :**61**(4): 456-502.

Brown, J. R., C. J. Douady, M. J. Italia, W. E. Marshall and M. J. Stanhope (2001). Universal trees based on large combined protein sequence data sets. Nature Genetics **28**(3): 281-5.

Bryant, D. (2003) A classification of consensus methods for phylogenies. in Janowitz, M., Lapointe, F.-J., McMorris, F.R., Mirkin, B., Roberts, F.S. (eds) BioConsensus, DIMACS 163-184.

Bryant, D. and V. Moulton (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Molecular Biology and Evolution **21**(2): 255-65.

Bull, J. J., J. P. Huelsenbeck, C. W. Cunningham, D. L. Swofford and P. J. Waddell (1993). Partitioning and Combining Data in Phylogenetic Analysis. Systematic Biology **42**(3): 384-397.

Bush, R. M., W. M. Fitch, C. A. Bender and N. J. Cox (1999). Positive selection on the H3 hemagglutinin gene of human influenza virus A. Molecular Biology and Evolution **16**(11): 1457-65.

Bushman, F. (2001) Lateral DNA transfer: Mechanisms and consequences. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Byrd, D. R. and S. W. Matson (1997). Nicking by transesterification: the reaction catalysed by a relaxase. Molecular Microbiology **25**(6): 1011-22.

Cartwright, K. A. V. (1995). Meningococcal carriage and disease. Meningococcal disease: 177-205.

Castillo-Davis, C. I. and D. L. Hartl (2003). Conservation, relocation and duplication in genome evolution. Trends in Genetics **19**(11): 593-7.

Cavalli-Sforza, L.L. and Edwards, A.W. (1967) Phylogenetic analysis: models and estimation procedures, American Journal of Human Genetics, 19 233-257

Charleston, M. A. (1998). Spectrum: spectral analysis of phylogenetic data. Bioinformatics **14**(1): 98-9.

Chen, D., L. Diao, O. Eulenstein, D. Fernández-Baca and M. J. Sanderson (2003). "Flipping: A Supertree Construction Method." DIMACS Series in Discrete Mathematics and Theoretical Computer Science **61**: 135 - 161.

Claverys, J. P. and B. Martin (2003). Bacterial competence genes: signatures of active transformation, or only remnants? Trends in Microbiology **11**(4): 161-5.

Clyde, W. C., and Fisher, D. C. (1997). Comparing the fit of stratigraphic and morphologic data in phylogenetic analysis. Paleobiology 23:1-19

Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, 3rd, et al. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature **393**(6685): 537-44.

Creevey, C. J. and J. O. McInerney (2002). An algorithm for detecting directional and non-directional positive selection, neutrality and negative selection in protein coding DNA sequences. Gene **300**(1-2): 43-51.

Creevey, C. J. and J. O. McInerney (2003). CRANN: detecting adaptive evolution in protein-coding DNA sequences. Bioinformatics **19**(13): 1726-.

Creevey, C. J. and J. O. McInerney (2004). Clann: investigating phylogenetic information through supertree analyses. Bioinformatics (In press).

Creevey C. J., Fitzpatrick, D. A., Philip, G. K., Kinsella, R. J., O'Connell, M. J., Pentony, M. M., Travers, S. A., Wilkinson, M. and McInerney, J. O. (2004) Does a tree-like phylogeny only exist at the tips in the prokaryotes? (In press) Proceedings of the Royal Society B Series: Biological Sciences.

Cummings, L. M., S. A. Handley, D. S. Myers, D. L. Reed, A. Rokas and K. Winka (2003). Comparing Bootstrap and Posterior Probability Values in the Four-Taxon Case. Systematic Biology **52**(4): 477-487.

Darwin, C. (1859). On The Origin Of Species By Means Of Natural Selection. London, John Murray.

Daubin, V., M. Gouy and G. Perriere (2001). Bacterial molecular phylogeny using supertree approach. Genome Inform Ser Workshop Genome Inform. 2001;12:155-64.

Daubin, V., M. Gouy and G. Perriere (2002). A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Research **12**(7): 1080-90.

Daubin, V., E. Lerat and G. Perriere (2003). The source of laterally transferred genes in bacterial genomes. Genome Biology **4**(9): R57.

Davidsen, T., E. A. Rodland, K. Lagesen, E. Seeberg, T. Rognes and T. Tonjum (2004). Biased distribution of DNA uptake sequences towards genome maintenance genes. Nucleic Acids Res **32**(3): 1050-8.

Davison, J. (1999). Genetic exchange between bacteria in the environment. Plasmid **42**(2): 73-91.

De Queiroz, A. (1993). For consensus (sometimes). Systematic Biology **42**: 368-372.

Deckert, G., P. V. Warren, T. Gaasterland, W. G. Young, A. L. Lenox, D. E. Graham, R. Overbeek, M. A. Snead, M. Keller, M. Aujay, et al. (1998). The complete genome of the hyperthermophilic bacterium Aquifex aeolicus. Nature **392**(6674): 353-8.

Deich, R. A. and H. O. Smith (1980). Mechanism of homospecific DNA uptake in Haemophilus influenzae transformation. Molecular & General Genetics : **177**(3): 369-74.

Delcher, A. L., D. Harmon, S. Kasif, O. White and S. L. Salzberg (1999). Improved microbial gene identification with GLIMMER. Nucleic Acids Research **27**(23): 4636-41.

de la Cruz, F. and J. Davies (2000). Horizontal gene transfer and the origin of species: lessons from bacteria. Trends in Microbiology **8**(3): 128-33.

Doolittle, R. F. (1998). Microbial genomes opened up. Nature **392**(6674): 339-42.

Doolittle, W. F. (1999a). Lateral genomics. Trends in Cell Biology **9**(12): M5-8.

Doolittle, W. F. (1999b). Phylogenetic classification and the universal tree. Science **284**(5423): 2124-9.

Doolittle, W. F. (2000a). The nature of the universal ancestor and the evolution of the proteome. Current Opinion in Structural Biology **10**(3): 355-8.

Doolittle, W. F. (2000b). Uprooting the tree of life. Science America **282**(2): 90-5.

Dubnau, D. (1997). Binding and transport of transforming DNA by Bacillus subtilis: the role of type-IV pilin-like proteins-a review. Gene **192**(1): 191-8.

Dubnau, D. (1999). DNA uptake in bacteria. Annual Review of Microbiology **53**: 217-44.

Dykhuizen, D. E. and G. Baranton (2001). The implications of a low rate of horizontal transfer in Borrelia. Trends in Microbiology **9**(7): 344-50.

Eernisse, D. and A. Kluge (1993). Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. Molecular Biology and Evolution **10**(6): 1170-1195.

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Research **8**(3): 163-7.

Eisen, J. A. (2000). Assessing evolutionary relationships among microbes from whole-genome analysis. Current Opinion in Microbiology **3**(5): 475-80.

Elkins, C., C. E. Thomas, H. S. Seifert and P. F. Sparling (1991). Species-specific uptake of DNA by gonococci is mediated by a 10-base-pair sequence. Journal of Bacteriology **173**(12): 3911-3.

Emelyanov, V. V. and B. V. Sinitsyn. (1999). A groE-based phylogenetic analysis shows very close evolutionary relationship between mitochondria and Rickettsia. Russian Journal of Genetics 35:618-627.

Emelyanov, V. V. (2001). Evolutionary relationship of Rickettsiae and mitochondria. FEBS Letter **501**(1): 11-8.

Emelyanov, V. V. (2003). Mitochondrial connection to the origin of the eukaryotic cell. European Journal of Biochemistry **270**(8): 1599-618.

Endo, T., K. Ikeo and T. Gojobori (1996). Large-scale search for genes on which positive selection may operate. Molecular Biology and Evolution **13**(5): 685-90.

Erixon, P., B. Svennblad, T. Britton and B. Oxelman (2003). Reliability of Bayesian Posterior Probabilities and Bootstrap frequencies in Phylogenetics. Systematic Biology **52**(5): 665-673.

Esser, C., N. Ahmadinejad, C. Wiegand, C. Rotte, F. Sebastiani, G. Gelius-Dietrich, K. Henze, E. Kretschmann, E. Richly, D. Leister, et al. (2004). A Genome Phylogeny for Mitochondria Among {alpha}-Proteobacteria and a Predominantly Eubacterial Ancestry of Yeast Nuclear Genes. Molecular Biology and Evolution 21:1643-1660.

Faith, D. P. and P. S. Cranston (1991). Could a cladogram this short have arisen by chance alone? On permutation tests for cladistic structure. Cladistics **7**: 1-28.

Fares, M. A. (2004). SWAPSC: sliding window analysis procedure to detect selective constraints. Bioinformatics (In press).

Fares, M. A., S. F. Elena, J. Ortiz, A. Moya and E. Barrio (2002). A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. Journal of Molecular Evolution **55**(5): 509-21.

Fares, M. A., A. Moya, C. Escarmis, E. Baranowski, E. Domingo and E. Barrio (2001). Evidence for positive selection in the capsid protein-coding region of the foot-and-mouth disease virus (FMDV) subjected to experimental passage regimens. Molecular Biology and Evolution **18**(1): 10-21.

Fay, J. C., G. J. Wyckoff and C. I. Wu (2001). Positive and negative selection on the human genome. Genetics **158**(3): 1227-34.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783-791.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package. Cladistics **5**: 164-166.

Felsenstein, J. (2004). Inferring Phylogenies. Massachusetts, Sinauer Associates.

Fitch, W. M. (1997). Networks and viral evolution. Journal of Molecular Evolution **44**:65-75.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty and J. M. Merrick (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science **269**(5223): 496-512.

Flook, P. K., S. Klee and C. H. Rowell (1999). Combined molecular phylogenetic analysis of the Orthoptera (Arthropoda, Insecta) and implications for their higher systematics. Systematic Biology **48**(2): 233-53.

Frasch, C. E. (1989). Vaccines for prevention of meningococcal disease. Clinical Microbiology Reviews **2 Suppl**: S134-8.

Fraser, C. M., S. Casjens, W. M. Huang, G. G. Sutton, R. Clayton, R. Lathigra, O. White, K. A. Ketchum, R. Dodson, E. K. Hickey, et al. (1997). Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. Nature **390**(6660): 580-6.

Fraser, C. M., J. Eisen, R. D. Fleischmann, K. A. Ketchum and S. Peterson (2000). Comparative genomics and understanding of microbial biology. Emerging Infectious Diseases **6**(5): 505-12.

Fussenegger, M., T. Rudel, R. Barten, R. Ryll and T. F. Meyer (1997). Transformation competence and type-4 pilus biogenesis in Neisseria gonorrhoeae--a review. Gene **192**(1): 125-34.

Futterer, O., A. Angelov, H. Liesegang, G. Gottschalk, C. Schleper, B. Schepers, C. Dock, G. Antranikian and W. Liebl (2004). Genome sequence of Picrophilus torridus and its implications for life around pH 0. Proceedings of the National Academy of Sciences of the United States of America **101**(24): 9091-6.

Gamieldien, J., A. Ptitsyn and W. Hide (2002). Eukaryotic genes in Mycobacterium tuberculosis could have a role in pathogenesis and immunomodulation. Trends in Genetics **18**(1): 5-8.

Genevrois, S., L. Steeghs, P. Roholl, J. J. Letesson and P. van der Ley (2003). The Omp85 protein of Neisseria meningitidis is required for lipid export to the outer membrane. The Embo Journal **22**(8): 1780-9.

Gentle, I., K. Gabriel, P. Beech, R. Waller and T. Lithgow (2004). The Omp85 family of proteins is essential for outer membrane biogenesis in mitochondria and bacteria. The Journal of Cell Biology **164**(1): 19-24.

Giron, J. A., A. G. Torres, E. Freer and J. B. Kaper (2002). The flagella of enteropathogenic Escherichia coli mediate adherence to epithelial cells. Molecular Microbiology **44**(2): 361-79.

Gogarten, J. P., W. F. Doolittle and J. G. Lawrence (2002). Prokaryotic evolution in light of gene transfer. Molecular Biology and Evolution **19**(12): 2226-38.

Goldman, N. and Z. Yang (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution **11**(5): 725-36.

Goldschneider, I., E. C. Gotschlich and M. S. Artenstein (1969). Human immunity to the meningococcus. I. The role of humoral antibodies. The Journal of Experimental Medicine **129**(6): 1307-26.

Gotschlich, E. C., T. Y. Liu and M. S. Artenstein (1969). Human immunity to the meningococcus. 3. Preparation and immunochemical properties of the group A, group B, and group C meningococcal polysaccharides. The Journal of Experimental Medicine **129**(6): 1349-65.

Gould, S. J. and R. C. Lewontin (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Procedings of the Royal Society of London Biological Sciences **205**(1161): 581-98.

Grandi, G. (2003). Rational antibacterial vaccine design through genomic technologies. International Journal For Parasitology **33**(5-6): 615-20.

Grassly, N. and E. Holmes (1997). A likelihood method for the detection of selection and recombination using nucleotide sequences. Molecular Biology and Evolution **14**(3): 239-247.

Gray, M. W., G. Burger and B. F. Lang (1999). Mitochondrial evolution. Science **283**(5407): 1476-81.

Griffith, F. (1923). The influence of immune serum on the biological properties of the pneumococci. Ministry of Health, London.

Gu, X. and W. H. Li (1996). Bias-corrected paralinear and LogDet distances and tests of molecular clocks and phylogenies under nonstationary nucleotide frequencies. Molecular Biology and Evolution **13**(10): 1375-83.

Gupta, R. S. (1995). Evolution of the chaperonin families (Hsp60, Hsp10 and Tcp-1) of proteins and the origin of eukaryotic cells. Molecular Microbiology **15**(1): 1-11.

Hacker, J., U. Hentschel and U. Dobrindt (2003). Prokaryotic chromosomes and disease. Science **301**(5634): 790-3.

Haldane, J. B. S. (1928) Possible Worlds. New York: Hugh & Bros.

Hancock, R. E. (1998). Resistance mechanisms in Pseudomonas aeruginosa and other nonfermentative gram-negative bacteria. Clin Infect Dis **27** :93-9.

Hansmann, S. and W. Martin (2000). Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. International Journal of Systematic and Evolutionary Microbiology **50**: 1655-63.

Hasegawa, M., H. Kishino and T. Yano (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution **22**(2): 160-74.

Hoelzer, M. A. and R. E. Michod (1991). DNA repair and the evolution of transformation in Bacillus subtilis. III. Sex with damaged DNA. Genetics **128**(2): 215-23.

Holmes, E. C., R. Urwin and M. C. Maiden (1999). The influence of recombination on the population structure and evolution of the human pathogen Neisseria meningitidis. Molecular Biology and Evolution **16**(6): 741-9.

Holmes, E. C., C. H. Woelk, R. Kassis and H. Bourhy (2002). Genetic constraints and the adaptive evolution of rabies virus in nature. Virology **292**(2): 247-57.

Howe, M. M. (1973). Transduction by bacteriophage MU-1. Virology **55**(1): 103-17.

Huelsenbeck, J. P., J. J. Bull, and C. W. Cunningham. (1996). Combining data in phylogenetic analysis. Trends in Ecology and Evolution 11(4):152-158.

Huelsenbeck, J. P. and F. Ronquist (2001). MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**(8): 754-5.

Hughes, A. L. and M. Nei (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**(6186): 167-70.

Hughes, A. L. and M. Nei (1989). Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proceedings of the National Academy of Sciences of the United States of America **86**(3): 958-62.

Hughes, A. L.(1999). Adaptive evolution of genes and genomes. Oxford University Press, Oxford

Ina, Y. (1995). New methods for estimating the numbers of synonymous and nonsynonymous substitutions. Journal of Molecular Evolution **40**(2): 190-226.

Jain, R., M. C. Rivera and J. A. Lake (1999). Horizontal gene transfer among genomes: the complexity hypothesis. Proceedings of the National Academy of Sciences of the United States of America **96**(7): 3801-6.

Jain, R., M. C. Rivera, J. E. Moore and J. A. Lake (2002). Horizontal gene transfer in microbial genome evolution. Theoretical Population Biology **61**(4): 489-95.

Jain, R., M. C. Rivera, J. E. Moore and J. A. Lake (2003). Horizontal gene transfer accelerates genome innovation and evolution. Molecular Biology and Evolution **20**(10): 1598-602.

Jiggins, F. M., G. D. Hurst and Z. Yang (2002). Host-Symbiont Conflicts: Positive Selection on an Outer Membrane Protein of Parasitic but not Mutualistic Rickettsiaceae. Molecular Biology and Evolution **19**(8): 1341-9.

Jones, K. E., A. Purvis, A. MacLarnon, O. R. Bininda-Emonds and N. B. Simmons (2002). A phylogenetic supertree of the bats (Mammalia: Chiroptera). Biological Reviews of the Cambridge Philosophical Society **77**(2): 223-59.

Jukes, T. H. and C. R. Cantor. (1969). Evolution of protein molecules. In H. N. Munro, ed., Mammalian Protein Metabolism, pp. 21-132, Academic Press, New York

Kahn, M. E. and H. O. Smith (1984). Transformation in Haemophilus: a problem in membrane biology. The Journal of Membrane Biology **81**(2): 89-103.

Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno and M. Hattori (2004). The KEGG resource for deciphering the genome. Nucleic Acids Res **32:** 277-80.

Karlin, S. and L. Brocchieri (2000). Heat shock protein 60 sequence comparisons: duplications, lateral transfer, and mitochondrial evolution. Proceedings of the National Academy of Sciences of the United States of America **97**(21): 11348-53.

Kimura, M. (1968). Evolutionary rate at the molecular level. Nature **217**(129): 624-6.

Kimura, M. (1979). The neutral theory of molecular evolution. Scientific American **241**(5): 98-100.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution **16**(2): 111-20.

Kimura, M. (1983). The neutral theory of molecular evolution. Cambridge, Cambridge University Press.

King, J. L. and T. H. Jukes (1969). Non-Darwinian evolution. Science **164**(881): 788-98.

King, L. M. (1998). The role of gene conversion in determining sequence variation and divergence in the Est-5 gene family in Drosophila pseudoobscura. Genetics **148**(1): 305-15.

Kinsella, R. J., D. A. Fitzpatrick, C. J. Creevey and J. O. McInerney (2003). Fatty acid biosynthesis in Mycobacterium tuberculosis: lateral gene transfer, adaptive evolution, and gene duplication. Proceedings of the National Academy of Sciences of the United States of America **100**(18): 10320-5.

Kinsella, R. J. and J. O. McInerney (2003). Eukaryotic genes in Mycobacterium tuberculosis? Possible alternative explanations. Trends in Genetics **19**(12): 687-9.

Kleinschmidt, J. H. (2003). Membrane protein folding on the example of outer membrane protein A of Escherichia coli. Cellular and Molecular Life Sciences : **60**(8): 1547-58.

Klenk, H. P., R. A. Clayton, J. F. Tomb, O. White, K. E. Nelson, K. A. Ketchum, R. J. Dodson, M. Gwinn, E. K. Hickey, J. D. Peterson, et al. (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus. Nature **390**(6658): 364-70.

Kluge, A. G. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Systematic Biology **38**: 7-25.

Kluge, A. G. (1993). "3-taxon transformation in phylogenetic inference – ambiguity and distortion as regards explanatory power." Cladistics **9**(2): 246-259.

Koomey, M. (1998). Competence for natural transformation in Neisseria gonorrhoeae: a model system for studies of horizontal gene transfer. APMIS Supplement **84**: 56-61.

Korbel, J. O., B. Snel, M. A. Huynen and P. Bork (2002). SHOT: a web server for the construction of genome phylogenies. Trends in Genetics **18**(3): 158-62.

Kreitman, M. and M. Aguade (1986). Genetic uniformity in two populations of Drosophila melanogaster as revealed by filter hybridization of four-nucleotide-recognizing restriction enzyme digests. Proceedings of the National Academy of Sciences of the United States of America **83**(10): 3562-6.

Kurland, C. G. and S. G. Andersson (2000). Origin and evolution of the mitochondrial proteome. Microbiology and Molecular Biology Reviews **64**(4): 786-820.

Kurland, C. G., B. Canback and O. G. Berg (2003). Horizontal gene transfer: a critical view. Proceedings of the National Academy of Sciences of the United States of America **100**(17): 9658-62.

Lanave, C., G. Preparata, C. Saccone and G. Serio (1984). A new method for calculating evolutionary substitution rates. Journal of Molecular Evolution **20**(1): 86-93.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. (2001). Initial sequencing and analysis of the human genome. Nature **409**(6822): 860-921.

Lang, B. F., M. W. Gray and G. Burger (1999). Mitochondrial genome evolution and the origin of eukaryotes. Annual Review of Genetics **33**: 351-97.

Lanka, E. and B. M. Wilkins (1995). DNA processing reactions in bacterial conjugation. Annual Review of Biochemistry **64**: 141-69.

Lanyon, S. M. (1993). Phylogenetic frameworks: Towards a firmer foundation for the comparative approach. Biological journal of the Linnean Society **49**: 45-61.

Lapointe J.F. and G. Cucumel, (1997) The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa, Systematic. Biology. 46:2, 306-312.

Lawrence, J. G. and H. Ochman (1997). Amelioration of bacterial genomes: rates of change and exchange. Journal of Molecular Evolution **44**(4): 383-97.

Lawrence, J. G. and H. Ochman (1998). Molecular archaeology of the Escherichia coli genome. Proceedings of the National Academy of Sciences of the United States of America **95**(16): 9413-7.

Lee, Y. H., T. Ota and V. D. Vacquier (1995). Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Molecular Biology and Evolution **12**(2): 231-8.

Lento, G. M., R. E. Hickson, G. K. Chambers and D. Penny (1995). Use of spectral analysis to test hypotheses on the origin of pinnipeds. Molecular Biology and Evolution **12**(1): 28-52.

Levasseur, C. and F. J. Lapointe (2001). War and peace in phylogenetics: a rejoinder on total evidence and consensus. Systematic Biology **50**(6): 881-91.

Li, W. H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. Journal of Molecular Evolution **36**(1): 96-9.

Liberles, D. A., D. R. Schreiber, S. Govindarajan, S. G. Chamberlin and S. A. Benner (2001). The adaptive evolution database (taed). Genome Biology **2**(4): 0003.1-0003.18

Liberles, D. A. and M. L. Wayne (2002). Tracking adaptive evolutionary events in genomic sequences. Genome Biology **3**(6): reviews1018.1–reviews1018.4

Lockhart, P., M. Steel, M. Hendy and D. Penny (1994). Recovering Evolutionary Trees under a More Realistic Model of Sequence. Molecular Biology and Evolution **11**(4): 605-612.

Loferer, I. I., I. I. Jacobi, I. I. Posch, I. I. Gauss, I. I. Meier-Ewert and I. I. Seizinger (2000). Integrated bacterial genomics for the discovery of novel antimicrobials. Drug Discovery Today **5**(3): 107-114.

Lorenz, M. G. and W. Wackernagel (1994). Bacterial gene transfer by natural genetic transformation in the environment. Microbiological Reviews **58**(3): 563-602.

Love, P. E., M. J. Lyle and R. E. Yasbin (1985). DNA-damage-inducible (din) loci are transcriptionally activated in competent Bacillus subtilis. Proceedings of the National Academy of Sciences of the United States of America **82**(18): 6201-5.

Lynn, D. J., A. T. Lloyd, M. A. Fares and C. O Farrelly (2004). Evidence of positively selected sites in mammalian alpha-defensins. Molecular Biology and Evolution **21**(5): 819-27.

Maiden, M. C. and I. M. Feavers (1995). Population genetics and global epidemiology of the human pathogen *Neisseria meningitidis*. Population genetics of bacteria. Cambridge University Press, Cambridge, England

Maiden, M. C., B. Malorny and M. Achtman (1996). A global gene pool in the neisseriae. Molecular Microbiology **21**(6): 1297-8.

Manting, E. H. and A. J. Driessen (2000). Escherichia coli translocase: the unravelling of a molecular machine. Molecular Microbiology **37**(2): 226-38.

Martin, D., N. Cadieux, J. Hamel and B. R. Brodeur (1997). Highly conserved Neisseria meningitidis surface protein confers protection against experimental infection. The Journal of Experimental Medicine **185**(7): 1173-83.

Martin, K., G. Morlin, A. Smith, A. Nordyke, A. Eisenstark and M. Golomb (1998). The tryptophanase gene cluster of Haemophilus influenzae type b: evidence for horizontal gene transfer. Journal of Bacteriology **180**(1): 107-18.

Margush, T. and F. R. McMorris. (1981). Consensus n-trees. Bulletin of Mathematical Biology 43: 239-244.

Maurice, M. M., D. S. Gould, J. Carroll, Y. Vugmeyster and H. L. Ploegh (2001). Positive selection of an MHC class-I restricted TCR in the absence of classical MHC class I molecules. Proceedings of the National Academy of Sciences of the United States of America **98**(13): 7437-42.

Maynard-Smith, J. and N. H. Smith (1998). Detecting recombination from gene trees. Molecular Biology and Evolution **15**(5): 590-9.

Mendel, G. (1865). Versuche über Pflanzen-Hybriden, Verhandlungen des naturforschenden Vereines in Brünn.

McDonald, J. H. and M. Kreitman (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature **351**(6328): 652-4.

McInerney, J. O. (1998). GCUA: general codon usage analysis. Bioinformatics **14**(4): 372-3.

Messier, W. and C. B. Stewart (1997). Episodic adaptive evolution of primate lysozymes. Nature **385**(6612): 151-4.

Mise, K. and R. Nakaya (1977). Transduction of R plasmids by bacteriophages P1 and P22: distinction between generalized and specialized transduction. Molecular & General Genetics : **157**(2): 131-8.

Mitton, J. B. (1997). Selection in natural populations. Oxford University Press.

Miyamoto, M. M. and W. M. Fitch (1995). Testing species phylogenies and phylogenetic methods with congruence. Systematic Biology **44**(1): 64-76.

Moriyama, E. N. and J. R. Powell (1997). Synonymous substitution rates in Drosophila: mitochondrial versus nuclear genes. Journal of Molecular Evolution **45**(4): 378-91.

Munoz, R., E. Garcia and R. Lopez (1998). Evidence for horizontal transfer from Streptococcus to Escherichia coli of the kfiD gene encoding the K5-specific UDP-glucose dehydrogenase. Journal of Molecular Evolution **46**(4): 432-6.

Naess, A., A. Halstensen, H. Nyland, S. H. Pedersen, P. Moller, R. Borgmann, J. L. Larsen and E. Haga (1994). Sequelae one year after meningococcal disease. Acta Neurologica Scandinavica **89**(2): 139-42.

Nassif, X. (2002). Genomics of Neisseria meningitidis. International journal of medical microbiology **291**(6-7): 419-23.

Nassif, X., C. Pujol, P. Morand and E. Eugene (1999). Interactions of pathogenic Neisseria with host cells. Is it possible to assemble the puzzle? Molecular Microbiology **32**(6): 1124-32.

Nei, M. and T. Gojobori (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution **3**(5): 418-26.

Nelson, K. E. (2003). The future of microbial genomics. Environmental Microbiology **5**(12): 1223-5.

Nesbo, C. L., Y. Boucher and W. F. Doolittle (2001). Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. Journal of Molecular Evolution **53**(4-5): 340-50.

Nichols, B. P. and C. Yanofsky (1979). Nucleotide sequences of trpA of Salmonella typhimurium and Escherichia coli: an evolutionary comparison. Proceedings of the National Academy of Sciences of the United States of America **76**(10): 5244-8.

Nielsen, R. and J. P. Huelsenbeck (2002). Detecting positively selected amino acid sites using posterior predictive P-values. Pacific Symposium for Biocomputing: 576-88.

Nowak, M. A., R. M. Anderson, A. R. McLean, T. F. Wolfs, J. Goudsmit and R. M. May (1991). Antigenic diversity thresholds and the development of AIDS. Science **254**(5034): 963-9.

Ochman, H., J. G. Lawrence and E. A. Groisman (2000). Lateral gene transfer and the nature of bacterial innovation. Nature **405**(6784): 299-304.

Ogata, H., S. Audic, P. Renesto Audiffren, P. E. Fournier, V. Barbe, D. Samson, V. Roux, P. Cossart, J. Weissenbach, J. M. Claverie, et al. (2001). Mechanisms of evolution in Rickettsia conorii and R. prowazekii. Science **293**(5537): 2093-8.

Ohta, T. (1993). Pattern of nucleotide substitutions in growth hormone-prolactin gene family: a paradigm for evolution by gene duplication. Genetics **134**(4): 1271-6.

Ohta, T. (1995). Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. Journal of Molecular Evolution **40**(1): 56-63.

Ohta, T. and M. Kimura (1971). Behavior of neutral mutants influenced by asociated overdominant loci in finite populations. Genetics **69**(2): 247-60.

Olsen, G. J., C. R. Woese and R. Overbeek (1994). The winds of (evolutionary) change: breathing new life into microbiology. Journal of Bacteriology **176**(1): 1-6.

Parkhill, J., M. Achtman, K. D. James, S. D. Bentley, C. Churcher, S. R. Klee, G. Morelli, D. Basham, D. Brown, T. Chillingworth, et al. (2000). Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491. Nature **404**(6777): 502-6.

Penny, D. and M. Hendy (1986). Estimating the reliability of evolutionary trees. Molecular Biology and Evolution **3**(5): 403-417.

Penny, D. and C. J. O'Kelly (1991). Eukaryote origins. Seeds of a universal tree. Nature **350**(6314): 106-7.

Perego, M. and J. A. Hoch (1996). Cell-cell communication regulates the effects of protein aspartate phosphatases on the phosphorelay controlling development in Bacillus subtilis. Proceedings of the National Academy of Sciences of the United States of America **93**(4): 1549-53.

Perry, A. C., I. J. Nicolson and J. R. Saunders (1988). Neisseria meningitidis C114 contains silent, truncated pilin genes that are homologous to Neisseria gonorrhoeae pil sequences. Journal Bacteriology **170**(4): 1691-7.

Philippe, H. and C. J. Douady (2003). Horizontal gene transfer and phylogenetics. Current Opinion in Microbiology **6**(5): 498-505.

Phillips, M. J., F. Delsuc and D. Penny (2004). Genome-Scale Phylogeny and the Detection of Systematic Biases. Molecular Biology and Evolution **21**(7): 1455-1458.

Piel, J., I. Hofer and D. Hui (2004). Evidence for a symbiosis island involved in horizontal acquisition of pederin biosynthetic capabilities by the bacterial symbiont of Paederus fuscipes beetles. Journal of Bacteriology **186**(5): 1280-6.

Pisani, D., A. M. Yates, M. C. Langer and M. J. Benton (2002). A genus-level supertree of the Dinosauria. Proceedings of the Royal Society of London. Series B: Biological Sciences **269**(1494): 915-21.

Pizza, M., V. Scarlato, V. Masignani, M. M. Giuliani, B. Arico, M. Comanducci, G. T. Jennings, L. Baldi, E. Bartolini, B. Capecchi, et al. (2000). Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science **287**(5459): 1816-20.

Posada, D. and K. A. Crandall (1998). MODELTEST: testing the model of DNA substitution. Bioinformatics **14**(9): 817-8.

Purvis, A. (1995). A composite estimate of primate phylogeny. Philosophical Transactions of the Royal Society London B Biological Sciences **348**(1326): 405-21.

Py, B., L. Loiseau and F. Barras (2001). An inner membrane platform in the type II secretion machinery of Gram-negative bacteria. EMBO Report **2**(3): 244-8.

Ragan, M. A. (1992). Matrix representation in reconstructing phylogenetic relationships among the eukaryotes. Biosystems **28**(1-3): 47-55.

Ragan, M. A. (2001). On surrogate methods for detecting lateral gene transfer. FEMS Microbiol Letters **201**(2): 187-91.

Ragan, M. A. (2001). On surrogate methods for detecting lateral gene transfer. Fems Microbiology Letters **201**(2): 187-91.

Rambaut, A. and N. C. Grassly (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Computer Applications in the Biosciences **13**: 235-238.

Read, T. D., R. C. Brunham, C. Shen, S. R. Gill, J. F. Heidelberg, O. White, E. K. Hickey, J. Peterson, T. Utterback, K. Berry, et al. (2000). Genome sequences of

Chlamydia trachomatis MoPn and Chlamydia pneumoniae AR39. Nucleic Acids Research **28**(6): 1397-406.

Redfield, R. J. (1993). Genes for breakfast: the have-your-cake-and-eat-it-too of bacterial transformation. Journal of Heredity **84**(5): 400-4.

Rohlf, F.J (1982). Consensus indices for comparing classifications. Mathematical Biosciences, 59:131-144.

Rokas, A., B. L. Williams, N. King and S. B. Carroll (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature **425**(6960): 798-804.

Ruta, M., J. E. Jeffery and M. I. Coates (2003). A supertree of early tetrapods. Proceedings of the Royal Society London Biological Sciences **270**(1532): 2507-16.

Saitou, N. and M. Nei (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution **4**(4): 406-25.

Salzberg, S. L., A. L. Delcher, S. Kasif and O. White (1998). Microbial gene identification using interpolated Markov models. Nucleic Acids Research **26**(2): 544-8.

Salzberg, S. L., O. White, J. Peterson and J. A. Eisen (2001). Microbial genes in the human genome: lateral transfer or gene loss? Science **292**(5523): 1903-6.

Sanderson, M. J. (1995). Objections to Bootstrapping Phylogenies: A Critique. Systematic Biology **44**(3): 299-320.

Sawyer, S. A. and D. L. Hartl (1992). Population genetics of polymorphism and divergence. Genetics **132**(4): 1161-76.

Schmalhausen, I. I. (1949). Factors of evolution. Blakiston, Philadelphia, Pennsylvania, USA.

Schmidt, H. A., K. Strimmer, M. Vingron and A. von Haeseler (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18**(3): 502-4.

Schuchat, A., K. Robinson, J. D. Wenger, L. H. Harrison, M. Farley, A. L. Reingold, L. Lefkowitz and B. A. Perkins (1997). Bacterial meningitis in the United States in 1995. Active Surveillance Team. The New England Journal of Medicine **337**(14): 970-6.

Sharp, P. M. (1997). In search of molecular darwinism. Nature **385**(6612): 111-2.

Shimada, K., R. A. Weisberg and M. E. Gottesman (1972). Prophage lambda at unusual chromosomal locations. I. Location of the secondary attachment sites and the properties of the lysogens. Journal Molecular Biology **63**(3): 483-503.

Shimada, K., R. A. Weisberg and M. E. Gottesman (1973). Prophage lambda at unusual chromosomal locations. II. Mutations induced by bacteriophage lambda in Escherichia coli K12. Journal Molecular Biology **80**(2): 297-314.

Shimodaira, H. and M. Hasegawa (1999). Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Molecular Biology and Evolution **16**(8): 1114-1116.

Shirai, M., H. Hirakawa, M. Kimoto, M. Tabuchi, F. Kishi, K. Ouchi, T. Shiba, K. Ishii, M. Hattori, S. Kuhara, et al. (2000). Comparison of whole genome sequences of Chlamydia pneumoniae J138 from Japan and CWL029 from USA. Nucleic Acids Research **28**(12): 2311-4.

Sicheritz-Ponten, T., C. G. Kurland and S. G. Andersson (1998). A phylogenetic analysis of the cytochrome b and cytochrome c oxidase I genes supports an origin of mitochondria from within the Rickettsiaceae. Biochim Biophys Acta **1365**(3): 545-51.

Sierra, G. V., H. C. Campa, N. M. Varcacel, I. L. Garcia, P. L. Izquierdo, P. F. Sotolongo, G. V. Casanueva, C. O. Rico, C. R. Rodriguez and M. H. Terry (1991). Vaccine against group B Neisseria meningitidis: protection trial and mass vaccination results in Cuba. Niph Annals **14**(2): 195-207.

Slowinski, J. B. and B. I. Crother (1998). Is the PTP test useful. Cladistics **14**: 297-302.

Smith, N. G. and A. Eyre-Walker (2002). Adaptive protein evolution in Drosophila. Nature **415**(6875): 1022-4.

Smith, N. H., J. Maynard-Smith and B. G. Spratt (1995). Sequence evolution of the porB gene of Neisseria gonorrhoeae and Neisseria meningitidis: evidence of positive Darwinian selection. Molecular Biology and Evolution **12**(3): 363-70.

Snel, B., P. Bork and M. A. Huynen (1999). Genome phylogeny based on gene content. Nature Genetics **21**(1): 108-10.

Snel, B., P. Bork and M. A. Huynen (2002). Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Research **12**(1): 17-25.

Sonea, S. and L. G. Mathieu (2001). Evolution of the genomic systems of prokaryotes and its momentous consequences. International Microbiology **4**(2): 67-71.

Stanhope, M. J., A. Lupas, M. J. Italia, K. K. Koretke, C. Volker and J. R. Brown (2001). Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature **411**(6840): 940-4.

Steel, M., A. W. Dress and S. Bocker (2000). Simple but fundamental limitations on supertree and consensus tree methods. Systematic Biology **49**(2): 363-8.

Stephens, R. S., S. Kalman, C. Lammel, J. Fan, R. Marathe, L. Aravind, W. Mitchell, L. Olinger, R. L. Tatusov, Q. Zhao, et al. (1998). Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. Science **282**(5389): 754-9.

Stephens, R. S. and C. J. Lammel (2001). Chlamydia outer membrane protein discovery using genomics. Current Opinion in Microbiology **4**(1): 16-20.

Suzuki, Y. (2004). Negative selection on neutralization epitopes of poliovirus surface proteins: implications for prediction of candidate epitopes for immunization. Gene **328**: 127-33.

Suzuki, Y., G. V. Glazko and M. Nei (2002). Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proceedings of the National Academy of Sciences of the United States of America **99**(25): 16138-16143.

Suzuki, Y. and T. Gojobori (1999). A method for detecting positive selection at single amino acid sites. Molecular Biology and Evolution **16**(10): 1315-1328.

Suzuki, Y. and T. Gojobori (1999). A method for detecting positive selection at single amino acid sites. Molecular Biology and Evolution **16**(10): 1315-28.

Suzuki, Y. and M. Nei (2001). Reliabilities of Parsimony-based and Likelihood-based Methods for Detecting Positive Selection at Single Amino Acid Sites. Molecular Biology and Evolution **18**(12): 2179-2185.

Suzuki, Y. and M. Nei (2001). Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. Molecular Biology and Evolution **18**(12): 2179-85.

Suzuki, Y. and M. Nei (2002). Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. Molecular Biology and Evolution **19**(11): 1865-9.

Suzuki, Y. and M. Nei (2004). False Positive Selection Identified by ML-Based Methods: Examples from the Sig1 Gene of the Diatom Thalassiosira weissflogii and the tax Gene of a Human T-Cell Lymphotropic Virus. Molecular Biology and Evolution: msh098.

Swanson, W. J., Z. Yang, M. F. Wolfner and C. F. Aquadro (2001). Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proceedings of the National Academy of Sciences of the United States of America **98**(5): 2509-14.

Swofford, D. L. (1998). PAUP*: Phylogenetic Analysis using Parsimony (*and Other Methods). Sinauer, Sunderland, MA.

Tamm, L. K., A. Arora and J. H. Kleinschmidt (2001). Structure and assembly of beta-barrel membrane proteins. The Journal of Biological Chemistry **276**(35): 32399-402.

Tamura, K. and S. Kumar (2002). Evolutionary distance estimation under heterogeneous substitution pattern among lineages. Molecular Biology and Evolution **19**(10): 1727-36.

Teifel, J. and H. Schmieger (1979). Phage Mu mutants with increased transduction abilities. Intervirology **11**(5): 314-6.

Tettelin, H., N. J. Saunders, J. Heidelberg, A. C. Jeffries, K. E. Nelson, J. A. Eisen, K. A. Ketchum, D. W. Hood, J. F. Peden, R. J. Dodson, et al. (2000). Complete genome sequence of Neisseria meningitidis serogroup B strain MC58. Science **287**(5459): 1809-15.

Thollesson, M. (2004). LDDist: a Perl module for calculating LogDet pair-wise distances for protein and nucleotide sequences. Bioinformatics **20**(3): 416-8.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research **22**(22): 4673-80.

Tsaur, S. C., C. T. Ting and C. I. Wu (1998). Positive selection driving the evolution of a gene of male reproduction, Acp26Aa, of Drosophila: II. Divergence versus polymorphism. Molecular Biology and Evolution **15**(8): 1040-6.

Urwin, R., E. C. Holmes, A. J. Fox, J. P. Derrick and M. C. Maiden (2002). Phylogenetic evidence for frequent positive selection and recombination in the meningococcal surface antigen PorB. Molecular Biology and Evolution **19**(10): 1686-94.

Voulhoux, R., M. P. Bos, J. Geurtsen, M. Mols and J. Tommassen (2003). Role of a highly conserved bacterial protein in outer membrane protein assembly. Science **299**(5604): 262-5.

Voulhoux, R. and J. Tommassen (2004). Omp85, an evolutionarily conserved bacterial protein involved in outer-membrane-protein assembly. Research in Microbiology **155**(3): 129-35.

Weinstock, G. M. (2000). Genomics and bacterial pathogenesis. Emerging Infectious Diseases **6**(5): 496-504.

Whelan, S. and N. Goldman (2001). A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Molecular Biology and Evolution **18**(5): 691-699.

Wiens, J. J. (1998). Combining Data Sets with Different Phylogenetic Histories. Systematic Biology **47**(4): 568-581.

Wilkinson M, Thorley JL, Littlewood DTJ, Bray RA. (2001). Towards a phylogenetic supertree of Platyhelminthes? In Interrelationships of the Platyhelminthes, ed. DTJ Littlewood, RA Bray, pp. 292-301. London: Taylor & Francis

Wilkinson, M. P. R., P. G. Peres Neto, P. G. Foster and C. B. Moncrieff (2002). Type 1 error rates of the parsimony permutation tail probability test. Systematic Biology **51**: 524-527.

Wilkinson, M., J. L. Thorley, D. Pisani, F. J. Lapointe and J. O. McInerney (2004). Some Desiderata for Liberal Supertrees. Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life. O. R. P. Bininda-Emonds. Boston, Kluwer Academic (In Press).

Winzer, K., Y. H. Sun, A. Green, M. Delory, D. Blackley, K. R. Hardie, T. J. Baldwin and C. M. Tang (2002). Role of Neisseria meningitidis luxS in cell-to-cell signaling and bacteremic infection. Infection and Immunity **70**(4): 2245-8.

Woese, C. R. (1987). Bacterial evolution. Microbiological Reviews **51**(2): 221-71.

Woese, C. R. (2000). Interpreting the universal phylogenetic tree. Proceedings of the National Academy of Sciences of the United States of America **97**(15): 8392-6.

Woese, C. R., O. Kandler and M. L. Wheelis (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences of the United States of America **87**(12): 4576-9.

Woese, C. R., E. Stackebrandt, T. J. Macke and G. E. Fox (1985). A phylogenetic definition of the major eubacterial taxa. Systematic and Applied Microbiology **6**: 143-51.

Wojciechowski, M. F., M. A. Hoelzer and R. E. Michod (1989). DNA repair and the evolution of transformation in Bacillus subtilis. II. Role of inducible repair. Genetics **121**(3): 411-22.

Wolf, Y. I., L. Aravind and E. V. Koonin (1999). Rickettsiae and Chlamydiae: evidence of horizontal gene transfer and gene exchange. Trends in Genetics **15**(5): 173-5.

Wolfgang, M., J. P. van Putten, S. F. Hayes, D. Dorward and M. Koomey (2000). Components and dynamics of fiber formation define a ubiquitous biogenesis pathway for bacterial pili. The Embo Journal **19**(23): 6408-18.

Woo, P. C., A. P. To, S. K. Lau and K. Y. Yuen (2003). Facilitation of horizontal transfer of antimicrobial resistance by transformation of antibiotic-induced cell-wall-deficient bacteria. Medical Hypotheses **61**(4): 503-8.

Wren, B. W. (2000). Microbial genome analysis: insights into virulence, host adaptation and evolution. Nature Review Genetics **1**(1): 30-9.

Wu, M., L. V. Sun, J. Vamathevan, M. Riegler, R. Deboy, J. C. Brownlie, E. A. McGraw, W. Martin, C. Esser, N. Ahmadinejad, et al. (2004). Phylogenomics of the Reproductive Parasite Wolbachia pipientis wMel: A Streamlined Genome Overrun by Mobile Genetic Elements. PLoS **2**(3): E69.

Xavier, K. B. and B. L. Bassler (2003). LuxS quorum sensing: more than just a numbers game. Current Opinion in Microbiology **6**(2): 191-7.

Xie, G., C. Forst, C. Bonner and R. A. Jensen (2002). Significance of two distinct types of tryptophan synthase beta chain in Bacteria, Archaea and higher plants. Genome Biol **3**(1):1-13.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution **39**(3): 306-14.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Computational Applied Biosciences **13**(5): 555-6.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in the Biosciences :**13**(5): 555-6.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution **15**(5): 568-73.

Yang, Z. (2000). Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. Journal of Molecular Evolution **51**(5): 423-32.

Yang, Z. (2002). Inference of selection from multiple species alignments. Current Opinion in Genetics & Development **12**(6): 688-94.

Yang, Z. and R. Nielsen (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular Biology and Evolution **17**(1): 32-43.

Yang, Z. and R. Nielsen (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular Biology and Evolution **19**(6): 908-17.

Yang, Z., R. Nielsen, N. Goldman and A. M. Pedersen (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**(1): 431-49.

Yang, Z. and W. J. Swanson (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Molecular Biology and Evolution **19**(1): 49-57.

Yang, Z. H. and J. P. Bielawski (2000). Statistical methods for detecting molecular adaptation. Trends in Ecology & Evolution **15**(12): 496-503.

Zanotto, P. M., E. G. Kallas, R. F. de Souza and E. C. Holmes (1999). Genealogical evidence for positive selection in the nef gene of HIV-1. Genetics **153**(3): 1077-89.

Zhang, J. (2004). Frequent false detection of positive selection by the likelihood method with branch-site models. Molecular Biology and Evolution **21**(7): 1332-9.

Zhang, J., Y. P. Zhang and H. F. Rosenberg (2002). Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nature Genetics **30**(4): 411-5.

Zharkikh, A. and W. H. Li (1992). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. Molecular Biology and Evolution **9**(6): 1119-47.

Publications

Fitzpatrick, D. A and McInerney, J. O. (2004) Evidence of positive Darwinian selection in Omp85, a highly conserved bacterial outer membrane protein essential for cell viability (In press) Journal of Molecular Evolution,

Creevey C. J., **Fitzpatrick, D. A**., Philip, G. K., Kinsella, R. J., O'Connell, M. J., Pentony, M. M., Travers, S. A., Wilkinson, M. and McInerney, J. O. (2004) Does a tree-like phylogeny only exist at the tips in the prokaryotes? (In press) Proceedings of the Royal Society B Series: Biological Sciences.

Kinsella, R. J., **D. A. Fitzpatrick**, C. J. Creevey and J. O. McInerney (2003). "Fatty acid biosynthesis in Mycobacterium tuberculosis: lateral gene transfer, adaptive evolution, and gene duplication." Proceedings of the National Academy of Sciences of the United States of America **100**(18): 10320-5.

Fitzpatrick, D. A, Creevey C. J. and McInerney, J. O. Evidence of positive Darwinian selection in putative meningococcal vaccine antigens.(Submitted)